EXHIBIT B

GENOMES

T.A. BROWN
Department of Biomolecular Sciences, UMIST, Manchester, M60 1QD, UK

WILEY-LISS
A JOHN WILEY & SONS, INC., PUBLICATION
New York • Chichester • Weinheim • Brisbane • Singapore • Toronto

Published in the United States of America, its dependent territories and Canada by John Wiley & Sons, Inc., by arrangement with BIOS Scientific Publishers Ltd, 9 Newtec Place, Magdalen Road, Oxford OX4 1RE, UK.

First published 1999

A CIP catalogue record for this book is available from the British Library.

ISBN 0–471–31618–0

Library of Congress Cataloging-in-Publication Data
Brown, T.A. (Terence A.)
Genomes / T.A. Brown.
p. cm.
Includes bibliographical references and index.
ISBN 0–471–31618–0 (pbk.)
1. Genomes. I. Title.
QH447.B76 1999
572.8'6--DC21 99–12241
CIP

USA
John Wiley & Sons Inc.,
605 Third Avenue, New York,
NY 10158–0012, USA

Canada
John Wiley & Sons (Canada) Ltd,
22 Worcester Road, Rexdale,
Ontario M9W 1L1, Canada

Production Editor: Fran Kingston
Typeset by J&L Composition Ltd, Filey, North Yorkshire, UK
Illustrations drawn by J&L Composition Ltd, Filey, North Yorkshire, UK
Printed by The Bath Press Ltd, Bath, UK

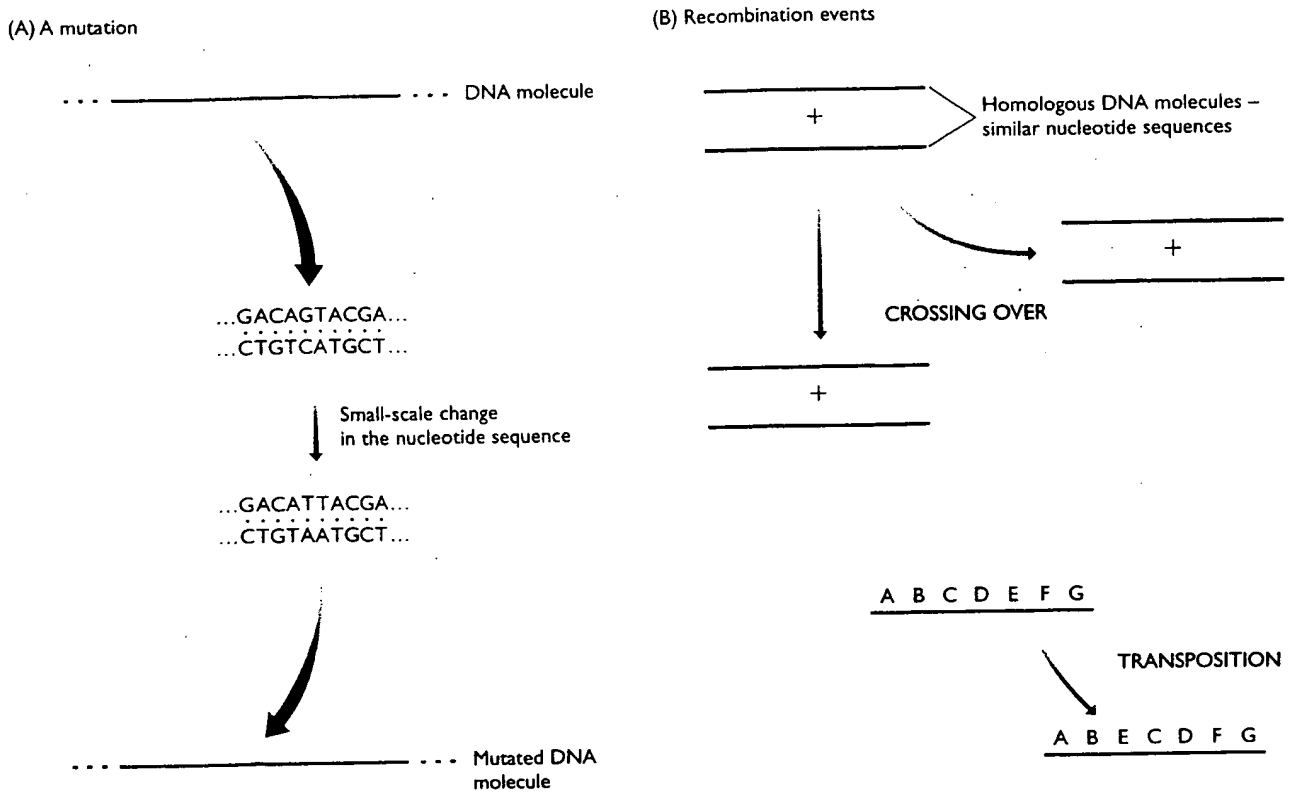


Figure 13.1 Mutation and recombination.

(A) A mutation is a small-scale change in the nucleotide sequence of a DNA molecule. A point mutation is shown but there are several other types of mutations, as described in the text. (B) Recombination events include exchange of segments of DNA molecules, as occurs during meiosis (see *Figure 2.10*, p. 26) and the movement of a segment from one position in a DNA molecule to another, for example by transposition (Section 13.2.3).

replication), is carried out and regulated by enzymes and other proteins.

Both mutation and recombination can have important effects on the cell in which they occur. A mutation in a key gene may cause the cell to die if it results in the protein coded by this gene being defective (Section 13.1.2) and some recombination events lead to changes in the biochemical capabilities of the cell, examples being those involved in immunoglobulin gene construction and yeast mating type switching. Other mutations and recombination events have a less significant impact on the phenotype of the cell and many have none at all. As we will see in Chapter 14, all mutations and recombination events that are not lethal have the potential to contribute to the evolution of the genome but for this to happen they must be inherited when the organism reproduces. With a single-celled organism such as a bacterium or yeast, all genome alterations that are not lethal or reversible are inherited by daughter cells and become permanent features of the lineage that descends for the original cell in which the alteration occurred. In a multicellular organism, only those events that occur in germ cells are relevant to genome evolution. Changes to the genomes of somatic cells are unimportant in an evolutionary sense, but they will have biological relevance if they result in a deleterious phenotype that affects the health of the organism.

13.1 MUTATIONS

With mutations, the issues that we have to consider are: how they arise; what effects they have on the genome and on the organism in which the genome resides; whether it is possible for a cell to increase its mutation rate and induce programmed mutations under certain circumstances; and how mutations are repaired.

13.1.1 The causes of mutations

Mutations arise in two ways:

- Some mutations are **spontaneous** errors in replication that evade the proofreading function of the DNA polymerases that synthesize new polynucleotides at the replication fork (Section 12.3.2). These mutations are called **mismatches** because they are

Box 13.1: Terminology for describing point mutations

Point mutations are also called simple mutations or single-site mutations. They are sometimes described as **substitution mutations** but this risks confusion because to an evolutionary geneticist 'substitution' occurs only when a mutation becomes fixed in a population (see Box 15.5, p. 408), so every individual displays it, as opposed to when the mutation first appears in a single organism.

Point mutations are divided into two categories:

- **Transitions** are purine-to-purine or pyrimidine-to-pyrimidine changes: A→G, G→A, C→T or T→C.
- **Transversions** are purine-to-pyrimidine or pyrimidine-to-purine changes: A→C, A→T, G→C, G→T, C→A, C→G, T→A or T→G.

positions where the nucleotide that is inserted into the daughter polynucleotide does not match, by base-pairing, the nucleotide at the corresponding position in the template DNA (*Figure 13.2A*). If the mismatch is retained in the daughter double helix then *one* of the granddaughter molecules produced during the next round of DNA replication will carry a permanent, double-stranded version of the mutation.

- Other mutations arise because a mutagen has reacted with the parent DNA, causing a structural change that affects the base-pairing capability of the altered nucleotide. Usually this alteration affects only one strand of the parent double helix, so only one of the daughter molecules carries the mutation, but two of the granddaughter molecules produced during the next round of replication will have it (*Figure 13.2B*).

Errors in replication are a source of point mutations

When considered purely as a chemical reaction, complementary base-pairing is not particularly accurate and if it was possible to copy a DNA template in the test tube, without the aid of any enzymes, then the resulting polynucleotide would probably have point mutations at 5–10 positions out of every hundred. This represents an error rate of 5–10%, which would be completely unacceptable during genome replication. The template-dependent DNA polymerases that carry out DNA replication must therefore increase the accuracy of the process by several orders of magnitude. This improvement is brought about in two ways:

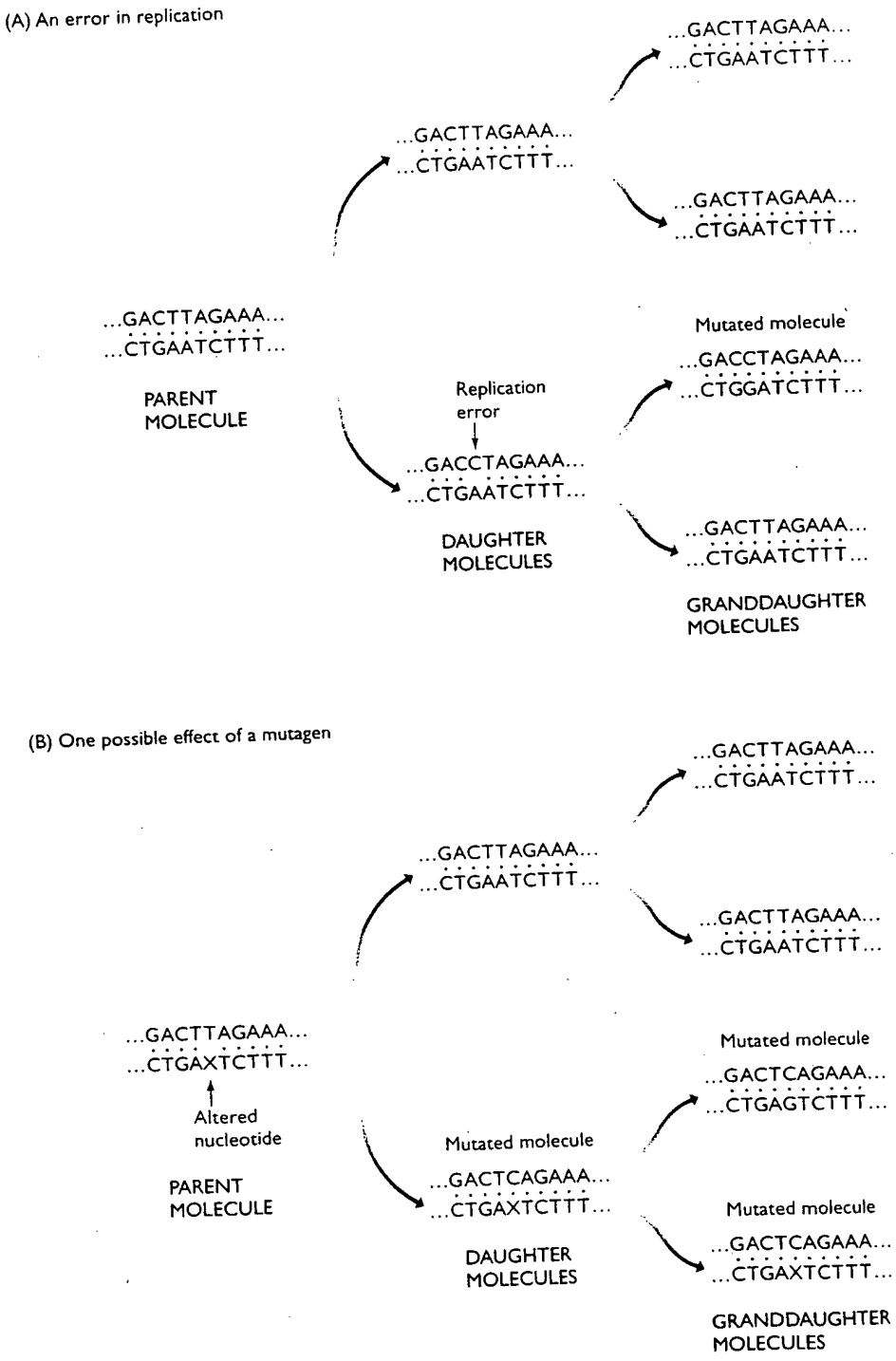
- The DNA polymerase operates a nucleotide selection process that dramatically increases the accuracy of template-dependent DNA synthesis (*Figure 13.3A*). This selection process probably acts at three different stages during the polymerization reaction, discrimination against an incorrect nucleotide occurring when the nucleotide is first bound to the DNA polymerase, when it is shifted to the active site of the enzyme, and when it is attached to the 3′-end of the polynucleotide that is being synthesized.

The accuracy of DNA synthesis is increased still further if the DNA polymerase possesses a 3′→5′ exonuclease activity and so is able to remove an incorrect nucleotide that evades the base selection process and becomes attached to the 3′-end of the new polynucleotide (see *Figure 12.11B*, p. 313). This is called **proofreading** (Section 12.3.2), but the name is a misnomer because the process is not an active checking mechanism. Instead, each step in the synthesis of a polynucleotide should be viewed as a competition between the polymerase and exonuclease functions of the enzyme, the polymerase usually winning because it is more active than the exonuclease, at least when the 3′-terminal nucleotide is base-paired to the template. But the polymerase activity is less efficient if the terminal nucleotide is not base-paired, the resulting pause in polymerization allowing the exonuclease activity to predominate so the incorrect nucleotide is removed (see *Figure 13.3B*).

Escherichia coli is able to synthesize DNA with an error rate of only 1 per 10^7 nucleotide additions. Interestingly, these errors are not evenly distributed between the two daughter molecules, the product of lagging strand replication being prone to about 20 times as many errors as the leading strand replicant. This asymmetry might indicate that DNA polymerase I, which is involved only in lagging strand replication (Section 12.3.2), has a less effective base selection and proofreading capability compared with DNA polymerase III, the main replicating enzyme (Francino and Ochman, 1997).

Not all of the errors that occur during DNA synthesis can be blamed on the polymerase enzymes: sometimes an error occurs even though the enzyme adds the 'correct' nucleotide, the one that base-pairs with the template. This is because each nucleotide base can occur as either of two alternative **tautomers**, structural isomers that are in dynamic equilibrium. For example, thymine exists as two tautomers, the *keto* and *enol* forms, with individual molecules occasionally undergoing a shift from one tautomer to the other. The equilibrium is biased very much towards the *keto* form but every now and then the *enol* version of thymine occurs in the template DNA at the precise time that the replication fork is moving past. This will lead to an 'error', because *enol*-thymine base-pairs with G rather than A (*Figure 13.4*). The same problem can occur with adenine, the rare *imino* tautomer of this base preferentially forming a pair with C, and guanine, *enol*-guanine pairing with thymine. After replication, the rare tautomer will inevitably revert to its more common form, leading to a mismatch in the daughter double helix.

As stated above, the error rate for DNA synthesis in *E. coli* is 1 in 10^7. The overall error rate for replication of the *E. coli* genome is only 1 in 10^{10} to 1 in 10^{11}, the



Examples of mutations.

(A) An error in replication leads to a mismatch in one of the daughter double helices, in this case a T to C change because one of the As in the template DNA was miscopied. When the mismatched molecule is itself replicated it gives one double helix with the correct sequence and one with a mutated sequence. (B) A mutagen has altered the structure of an A in the lower strand of the parent molecule, giving nucleotide X, which does not base-pair with the T in the other strand so, in effect, a mismatch has been created. When the parent molecule is replicated, X base-pairs with C, giving a mutated daughter molecule. When this daughter molecule is replicated, both granddaughters inherit the mutation.

TECHNICAL 13.1 NOTES

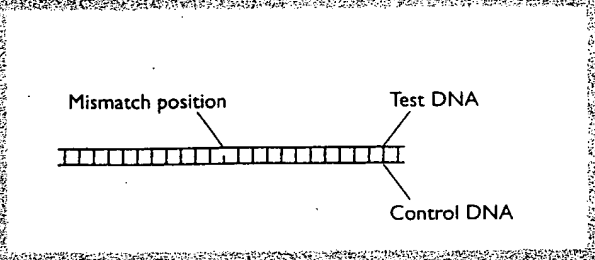
Mutation detection

Rapid procedures for detecting mutations in DNA molecules.

Many genetic diseases are caused by point mutations that result in modification or inactivation of a gene product. Methods for detecting these mutations are important in two contexts. First, when a new gene responsible for a genetic disease is first identified it is usually necessary to examine many versions of that gene, from different individuals, to identify the mutation or mutations responsible for the disease state. Second, when a disease-causing mutation has been characterized, high-throughput methods are needed so that clinicians can screen many DNA samples in order to identify individuals who have the mutation and are at risk of developing the disease or passing it on to their children.

Any mutation can be identified by DNA sequencing but sequencing is relatively slow and would be inappropriate for screening a large number of samples. DNA chip technology (Technical Note 2.3, p. 23) could also be employed, but this is not yet a widely available option. For these reasons, a number of 'low technology' methods have been devised. These can be divided into two categories: **mutation scanning** techniques, which require no prior information about the position of a mutation, and **mutation screening** techniques, which determine whether a specific mutation is present.

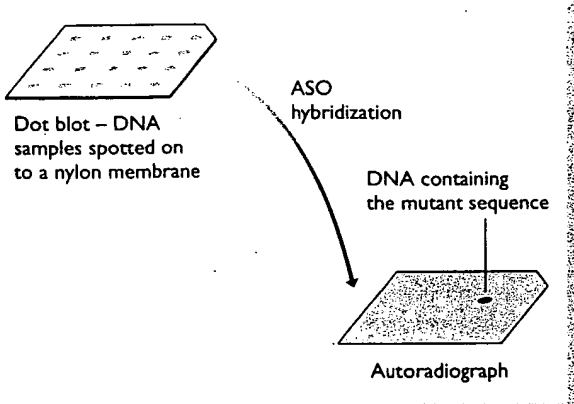
Most *scanning* techniques involve analysis of the heteroduplex formed between a single strand of the DNA being examined and the complementary strand of a control DNA that has the unmutated sequence:



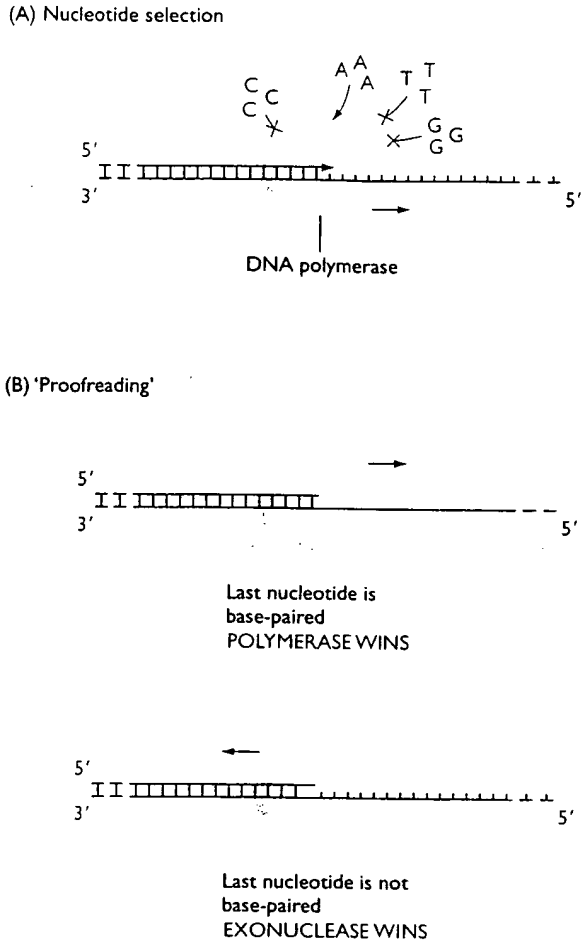
If the test DNA contains a mutation then there will be a single mismatched position in the heteroduplex, where a base pair has not formed. Various techniques can be used for detecting whether this mismatch is present or not (Cotton, 1997):

- **Electrophoresis** or **high-performance liquid chromatography (HPLC)** can detect the mismatch by identifying the difference in the mobility of the mismatched hybrid, compared with a fully base-paired one, in a polyacrylamide gel or HPLC column. This approach determines if a mismatch is present but does not provide information on where in the test DNA the mutation is located.
- **Cleavage** of the heteroduplex at the mismatch position followed by gel electrophoresis will locate the position of a mismatch. If the heteroduplex stays intact then no mismatch is present; if it is cleaved then it contains a mismatch, the position of the mutation in the test DNA being indicated by the sizes of the cleavage products. Cleavage is carried out by treatment with enzymes or chemicals that cut at single-stranded regions of mainly double-stranded DNA, or with a single-strand-specific ribonuclease such as S1 (see *Figure 5.7*, p. 94) if the hybrid has been formed between the control DNA and an RNA version of the test DNA.

Most *screening* methods for detection of specific mutations make use of the ability of oligonucleotides to distinguish between target DNAs whose sequences differ at just one nucleotide position (see *Figure 2.7*, p. 22). In **allele-specific oligonucleotide (ASO)** hybridization the DNA samples are screened by probing with an oligonucleotide that hybridizes only to the mutant sequence:



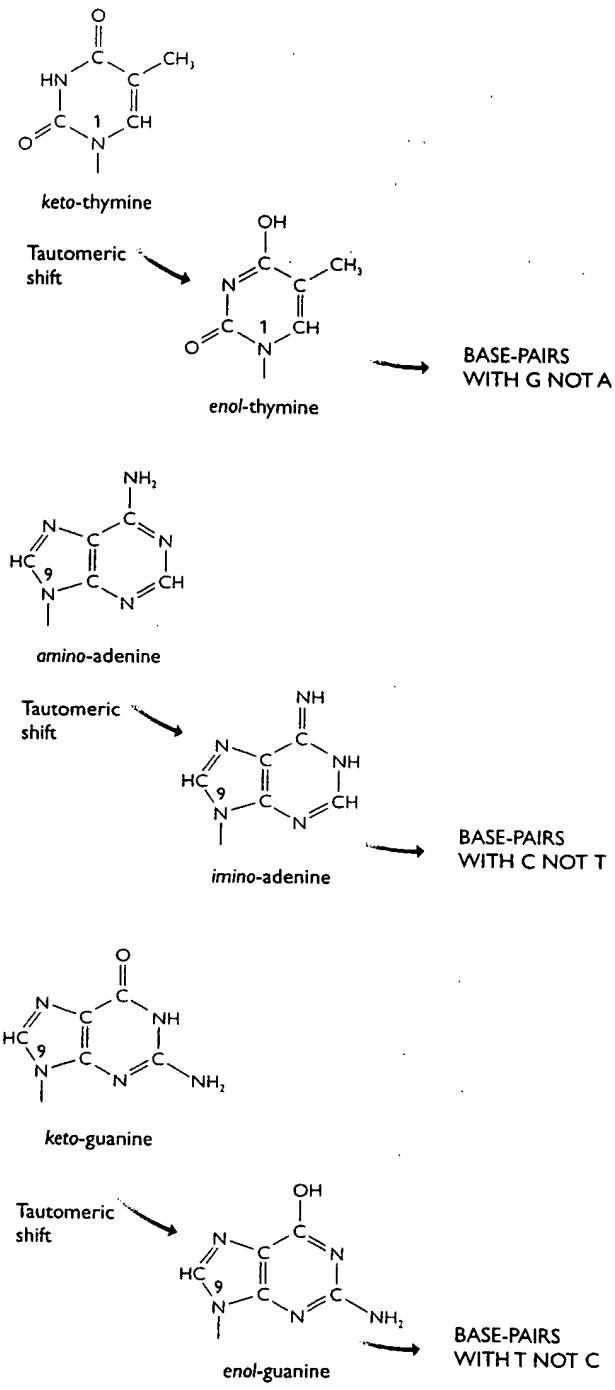
This is an efficient procedure but it is unnecessarily long-winded. The DNA samples are usually obtained by PCR of clinical isolates so a more rapid alternative is to use the diagnostic oligonucleotide as one of the PCR primers, so that presence or absence of the mutation in the test DNA is indicated by the synthesis or otherwise of a PCR product.



Mechanisms for ensuring the accuracy of DNA replication.

(A) The DNA polymerase actively selects the correct nucleotide to insert at each position. (B) Those errors that occur can be corrected by 'proofreading' if the polymerase has a 3′→5′ exonuclease activity. If the last nucleotide that was inserted is base-paired to the template then the polymerase activity predominates, but if the last nucleotide is not base-paired then the exonuclease activity is favored.

improvement compared to the polymerase error rate being due to the mismatch repair system (Section 13.1.4) that scans newly replicated DNA for positions where the bases are unpaired and hence corrects the few mistakes that the replication enzymes makes. The implication is that only one uncorrected replication error occurs every 1000 times that the *E. coli* genome is copied.



The effects of tautomerism on base-pairing.

In each of these three examples, the two tautomeric forms of the base have different pairing properties. Cytosine also has *amino* and *imino* tautomers but both pair with G.

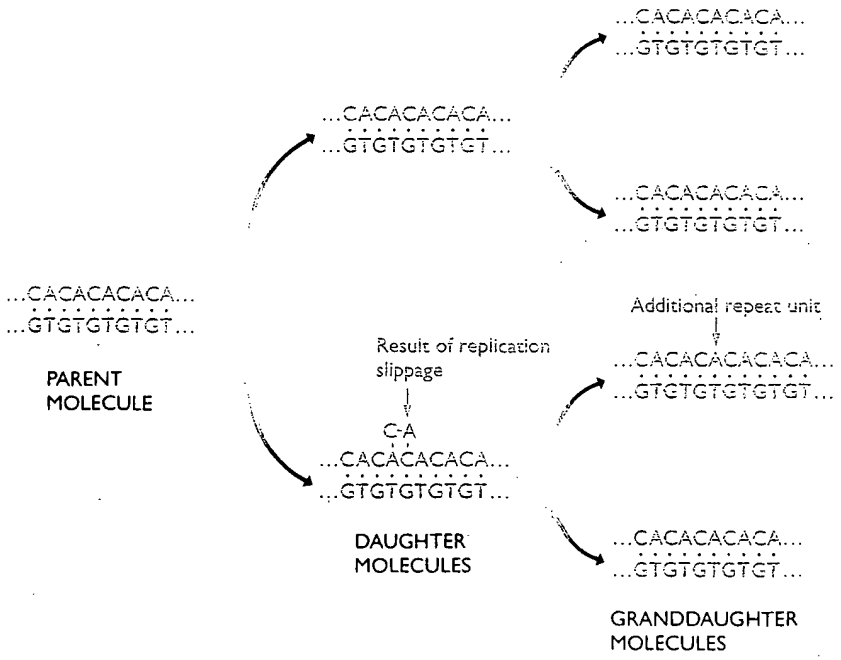
Not all errors in replication are point mutations. Aberrant replication can also result in small numbers of extra nucleotides being inserted into the polynucleotide being synthesized, or some nucleotides in the template not being copied. Insertions and deletions are often called **frameshift** mutations because when one occurs within a coding region it can result in a shift in the reading frame used for translation of the protein specified by the gene (see *Figure 13.12*, p. 342). However, it is inaccurate to use 'frameshift' to describe all insertions and deletions

because they can occur anywhere, not just in genes, and not all insertions or deletions in coding regions result in frameshifts: an insertion or deletion of three nucleotides, or multiples of three, simply adds or removes codons or parts of adjacent codons without affecting the reading frame.

Insertion and deletion mutations can affect all parts of the genome but are particularly prevalent when the template DNA contains short repeated sequences, such as are found in microsatellites (Section 6.3.1). This is because repeated sequences can induce **replication slippage**, in which the template strand and its copy shift their relative positions so that part of the template is either copied twice or missed out. The result is that the new polynucleotide has a larger or smaller number, respectively, of the repeat units (*Figure 13.5*). This is the main reason why microsatellite sequences are so variable, replication slippage occasionally generating a new length variant, adding to the collection of alleles already present in the population (Section 15.3.2).

Replication slippage is probably also responsible for the **trinucleotide repeat expansion diseases** that have been discovered in humans in recent years (Ashley and Warren, 1995). Each of these neurodegenerative diseases is due to a relatively short series of trinucleotide repeats becoming elongated to two or more times its normal length. For example, the human *Hdh* gene contains the sequence 5′-CAG-3′ repeated between 10 and 35 times in tandem, coding for a series of glutamines in the protein product. In Huntington's disease this repeat expands to a copy number of 36–121, increasing the length of the polyglutamine tract and resulting in a dysfunctional protein. Several other human diseases are also due to expansions of polyglutamine codons (*Table 13.1*). Some diseases associated with mental retardation result from trinucleotide expansions in the leader region of a gene, giving a **fragile site**, a position where the chromosome is likely to break (Sutherland *et al.*, 1998) and expansions involving intron and trailer regions are also known.

How triplet expansions are generated is not precisely understood. The size of the insertion is much greater than occurs with normal replication slippage, such as is seen with microsatellite sequences, and once the expansion reaches a certain length it appears to become susceptible to further expansion in subsequent rounds of replication, leading to the disease becoming increasingly severe in succeeding generations. The possibility that expansion involves formation of hairpin loops in the DNA has been raised, based on the observation that only a limited number of trinucleotide sequences are known to undergo expansion, and all of these sequences are GC-rich and so might form stable secondary structures. Studies of similar triplet expansions in yeast have shown that these are more prevalent when the *RAD27* gene is inactivated (Freudenreich *et al.*, 1998), an interesting observation as *RAD27* is the yeast version of the mammalian gene for



Replication slippage.

The diagram shows replication of a 5-unit CA repeat microsatellite. Slippage has occurred during replication of the parent molecule, inserting an additional repeat unit into the newly synthesized polynucleotide of one of the daughter molecules. When this daughter molecule replicates it gives a granddaughter molecule whose microsatellite array is one unit longer than that of the original parent.

Table 13.1 Examples of human trinucleotide repeat expansions

Locus	Repeat sequence: Normal	Repeat sequence: Mutated	Associated disease
Polyglutamine expansions (all in coding regions of genes)			
Hdh	$(CAG)_{10-35}$	$(CAG)_{36-121}$	Huntington's disease
AR	$(CAG)_{11-33}$	$(CAG)_{38-66}$	Spinal and bulbar muscular atrophy
B37	$(CAG)_{7-25}$	$(CAG)_{49-75}$	Dentatoribral-pallidoluysian atrophy
MJD1	$(CAG)_{12-37}$	$(CAG)_{61-84}$	Machado–Joseph disease
SCA1	$(CAG)_{6-39}$	$(CAG)_{41-81}$	Spinocerebellar ataxia type 1
Fragile site expansions (probably all in the untranslated leader regions of genes)			
FRAXA	$(CGG)_{6-52}$	$(CGG)_{60-1000}$	Fragile X syndrome
FRAXE	$(GCC)_{7-35}$	$(GCC)_{130-750}$	Fragile XE mental retardation
FRAXF	$(GCC)_{6-29}$	$(GCC)_{300-1000}$	None
FRA11B	$(CGG)_{11}$	$(CGG)_{80-1000}$	Predisposed to Jacobsen syndrome
FRA16A	$(CCG)_{16-49}$	$(CCG)_{1000-1900}$	None
Other expansions (positions described below)			
DMPK	$(CTG)_{5-37}$	$(CTG)_{50-3000}$	Myotonic dystrophy
FRDA	$(GAA)_{40-60}$	$(GAA)_{>200}$	Friedreich's ataxia

For more details see Ashley and Warren (1995). The DMPK and FRDA expansions are in the trailer and intron regions of their genes, respectively, and are thought to affect RNA processing (Ashley and Warren, 1995; Campuzano *et al.*, 1996). There are also a few disease-causing mutations that involve expansions of longer sequences, such as progressive myoclonus epilepsy caused by a $(CCCCGCCCCGCG)_{2-3}$ to $(CCCCGCCCCGCG)_{>12}$ expansion in the promoter region of the *EPM1* locus (Mandel, 1997).

FEN1, the protein involved in processing of Okazaki fragments (Section 12.3.2). This might indicate that a trinucleotide repeat expansion is caused by an aberration in lagging strand synthesis.

Mutations are also caused by chemical and physical mutagens

Many chemicals that occur naturally in the environment have mutagenic properties and these have been supplemented in recent years with other chemical mutagens that result from human industrial activity. Physical agents such as radiation are also mutagenic. Most organisms are exposed to greater or lesser amounts of these various mutagens and their genomes suffer damage as a result.

The definition of the term 'mutagen' is *a chemical or physical agent that causes mutations*. This definition is important because it distinguishes mutagens from other types of environmental agent that cause damage to cells in ways other than by causing mutations (*Table 13.2*). There are overlaps between these categories (for example, some mutagens are also carcinogens) but each type of agent has a distinct biological effect. The definition of mutagen also makes a distinction between true mutagens and other agents that damage DNA without causing mutations, for example by causing breaks in DNA molecules. This type of damage may block replication and cause the cell to die, but it is not a mutation in the strict sense of the term and the causative agents are therefore not mutagens.

Mutagens cause mutations in three different ways:

- Some act as **base analogs** and are mistakenly used as substrates when new DNA is synthesized at the replication fork.
- Some react directly with DNA, causing structural changes that lead to miscopying of the template strand when the DNA is replicated. These structural changes are diverse, as we will see when we look at individual mutagens.
- Some mutagens act indirectly on DNA. They do not themselves affect DNA structure, but instead cause the cell to synthesize chemicals such as peroxides that have the direct mutagenic effect.

The range of mutagens is so vast that it is difficult to devise an all-embracing classification. We will therefore restrict our study to the most common types. For chemical mutagens these are as follows:

- **Base analogs** are purine and pyrimidine bases that are similar enough to the standard bases to be

Table 13.2 Categories of environmental agent that cause damage to living cells

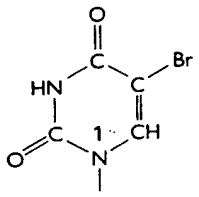
Agent	Effect on living cells
Carcinogen	Causes cancer – the neoplastic transformation of eukaryotic cells
Clastogen	Causes fragmentation of chromosomes
Mutagen	Causes mutations
Oncogen	Induces tumor formation
Teratogen	Results in developmental abnormalities

Based on Twyman (1998).

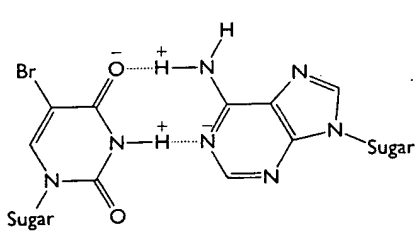
incorporated into nucleotides when these are synthesized by the cell. The resulting unusual nucleotides can then be used as substrates for DNA synthesis during genome replication. For example, **5-bromouracil** (**5-bU**; *Figure 13.6A*) has the same base-pairing properties as thymine and nucleotides containing this base can be added to the daughter polynucleotide at positions opposite A's in the template. The mutagenic effect arises because the equilibrium between the two tautomers of 5-bU is shifted more towards the rarer *enol* form than is the case with thymine. This means that during the next round of replication there is a relatively high chance of the polymerase encountering *enol*-5bU, which (like *enol*-thymine) pairs with G rather than A (*Figure 13.6B*). This results in a point mutation (*Figure 13.6C*). **2-Aminopurine** acts in a similar way: it is an analog of adenine with an *amino*-tautomer that pairs with thymine and an *imino*-tautomer that pairs with cytosine, the *imino* form being less uncommon than *imino*-adenine and hence inducing T to C transitions during DNA replication.

Deaminating agents also cause point mutations. A

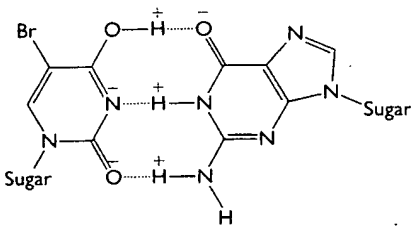
(A) 5-Bromouracil



(B) Base-pairing with 5-bromouracil



5-Bromouracil *keto* form — Adenine



5-Bromouracil *enol* form — Guanine

(C) The mutagenic effect of 5-bromouracil

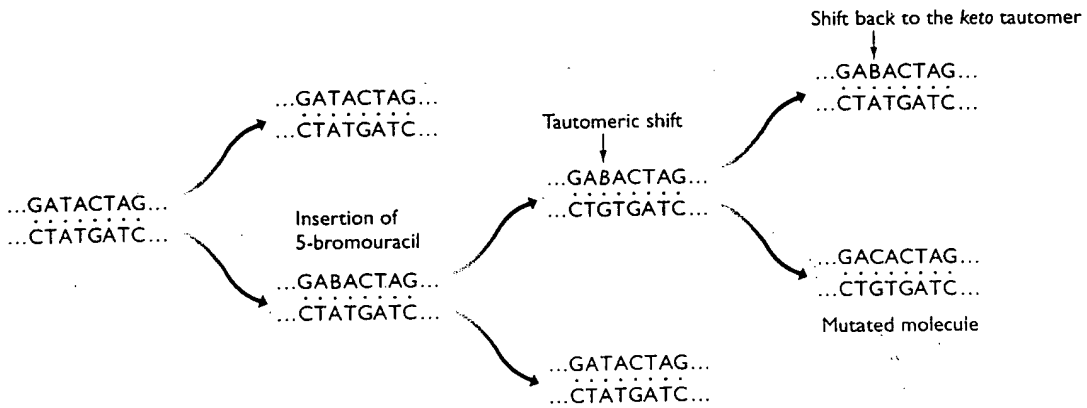


Figure 13.6 5-Bromouracil and its mutagenic effect.

See the text for details.

certain amount of base deamination (removal of an amino group) occurs spontaneously in genomic DNA molecules, with the rate being increased by chemicals such as nitrous acid, which deaminates adenine, cytosine and guanine (thymine has no amino group and so cannot be deaminated), and sodium bisulfite, which acts only on cytosine. Deamination of guanine is not mutagenic because the resulting base, xanthine, blocks replication when it appears in the template polynucleotide. Deamination of adenine gives hypoxanthine (*Figure 13.7*), which pairs with C rather than T, and deamination of cytosine gives uracil, which pairs with A rather than G. Deaminations of these two bases therefore result in point mutations when the template strand is copied.

- **Alkylating agents** are a third type of mutagen that can give rise to point mutations. Chemicals such as **ethylmethane sulfonate (EMS)** and dimethylnitrosamine add alkyl groups to nucleotides in DNA molecules, as do methylating agents such as methyl halides which are present in the atmosphere, and the products of nitrite metabolism. The effect that alkylation has depends on the position at which the nucleotide is modified and the type of alkyl group that is added. Methylations, for example, often result in modified nucleotides with altered base-pairing properties and so lead to point mutations. Other alkylations block replication by forming crosslinks between the two strands of a DNA molecule, or by adding large alkyl groups that prevent progress of the replication complex.
- **Intercalating agents** are usually associated with insertion mutations. The best known mutagen of this type is **ethidium bromide**, which fluoresces when exposed to UV radiation and so is used to reveal the positions of DNA bands after agarose gel electrophoresis (see Technical Note 3.2, p. 43). Ethidium bromide and other intercalating agents are flat molecules that can slip in between base pairs in the double helix, slightly unwinding the helix and hence increasing the distance between adjacent base pairs (*Figure 13.8*).

The most important types of physical mutagen are:

- **UV radiation** of 260 nm induces dimerization of adjacent pyrimidine bases, especially if these are both thymines (*Figure 13.9A*), resulting in a **cyclobutyl dimer**. Other pyrimidine combinations also form dimers, the order of frequency being 5′-CT-3′ > 5′-TC-3′ > 5′-CC-3′. Purine dimers are much less common. UV-induced dimerization usually results in a deletion mutation when the modified strand is copied. Another type of UV-induced **photoproduct** is the **(6–4) lesion** in which carbons number 4 and 6 of adjacent pyrimidines become covalently linked (*Figure 13.9B*).
- **Ionizing radiation** has various effects on DNA depending on the type of radiation and its intensity. Point, insertion and/or deletion mutations might arise, as well as more severe forms of DNA damage that prevent subsequent replication of the genome. Some types of ionizing radiation act directly on DNA, others act indirectly by stimulating the formation of reactive molecules such as peroxides in the cell.

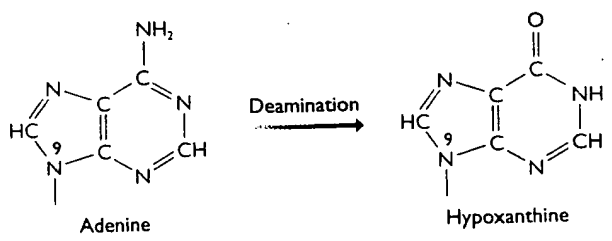


Figure 13.7 Hypoxanthine is a deaminated version of adenine.

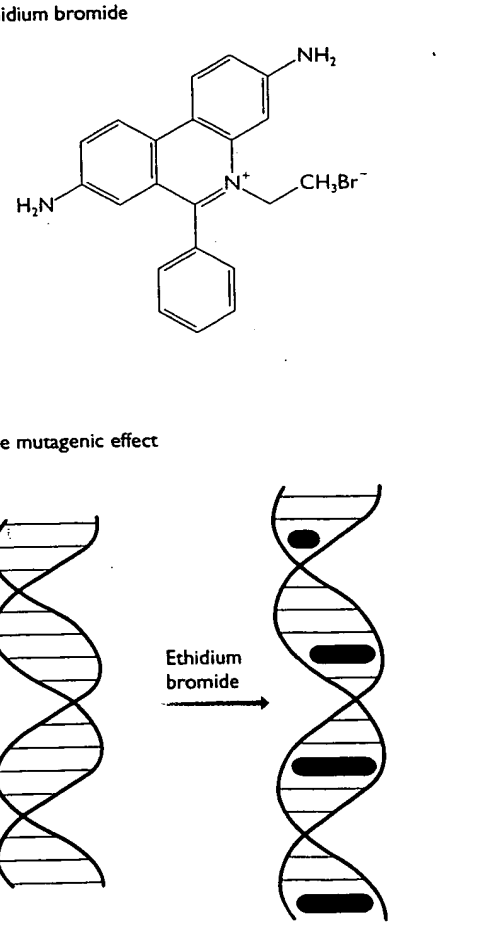
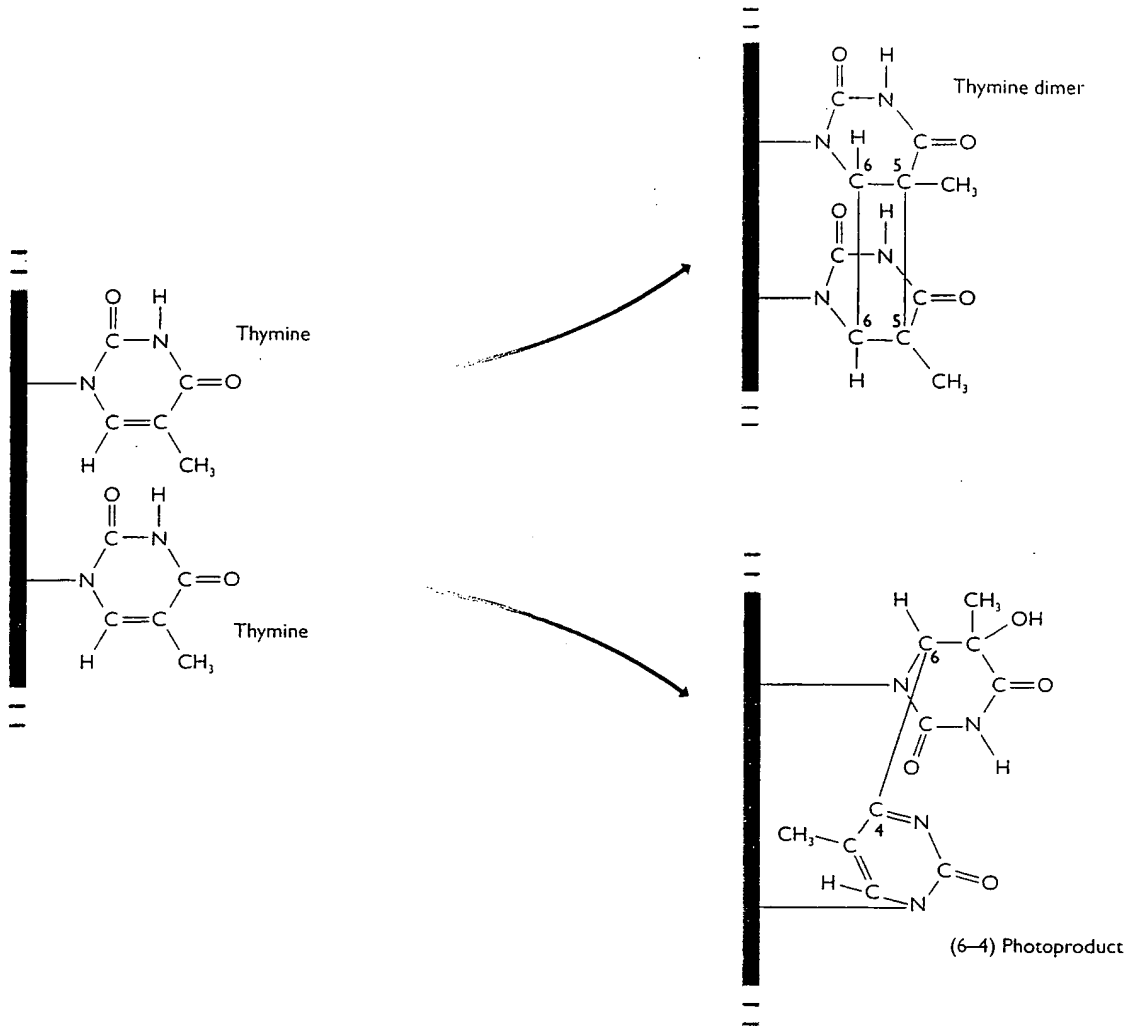


Figure 13.8 The mutagenic effect of ethidium bromide.

(A) Ethidium bromide is a flat, plate-like molecule that is able to slot in between the base pairs of the double helix. (B) Ethidium bromide molecules are seen intercalated into the helix: the molecules are viewed sideways on. Note that intercalation results in the distance between adjacent base pairs being increased.



Photoproducts induced by UV irradiation.

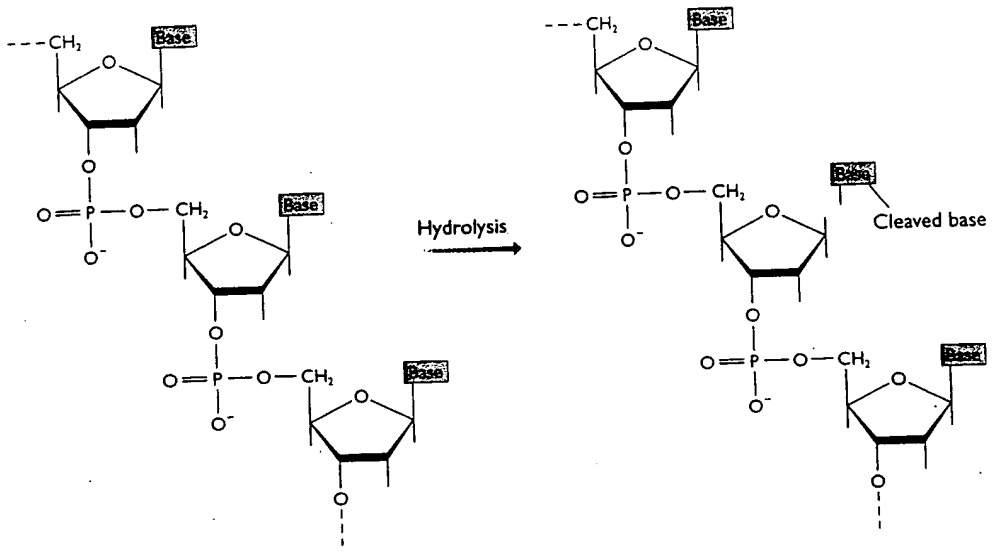
A segment of a polynucleotide containing two adjacent thymine bases is shown. (A) A thymine dimer contains two UV-induced covalent bonds, one linking the carbons at position 6 and the other linking the carbons at position 5. (B) The (6–4) lesion involves formation of a covalent bond between carbons 4 and 6 of the adjacent nucleotides.

Heat stimulates the water-induced cleavage of the β-*N*-glycosidic bond that attaches the base to the sugar component of the nucleotide (*Figure 13.10A*). This occurs more frequently with purines rather than pyrimidines and results in an **AP** (apurinic/apyrimidinic) or **baseless site**. The sugar–phosphate that is left is unstable and rapidly degrades, leaving a gap if the DNA molecule is double-stranded (*Figure 13.10B*). This reaction is not normally mutagenic because cells have effective systems for repairing nicks (Section 13.1.4), which is reassuring when one considers that 10 000 AP sites are generated in each human cell per day. Gaps do, however, lead to mutations under certain circumstances, for example in *E. coli* when the SOS response is activated, when gaps are filled with A's regardless of the identity of the nucleotide in the other strand (Section 13.1.3).

When considering the effects of mutations we must make a distinction between the *direct* effect that a mutation has on the functioning of a genome and its *indirect* effect on the phenotype of the organism in which it occurs. The direct effect is relatively easy to assess because we can use our understanding of gene structure and expression to predict the impact that a mutation will have on genome function. The indirect effects are more complex because these relate to the phenotype of the mutated organism which, as described in Section 5.2.2, is often difficult to correlate with the activities of individual genes.

Many mutations result in nucleotide sequence changes that have no effect on the functioning of the genome

(A) Heat-induced hydrolysis of a β-*N*-glycosidic bond



(B) The effect of hydrolysis on double-stranded DNA

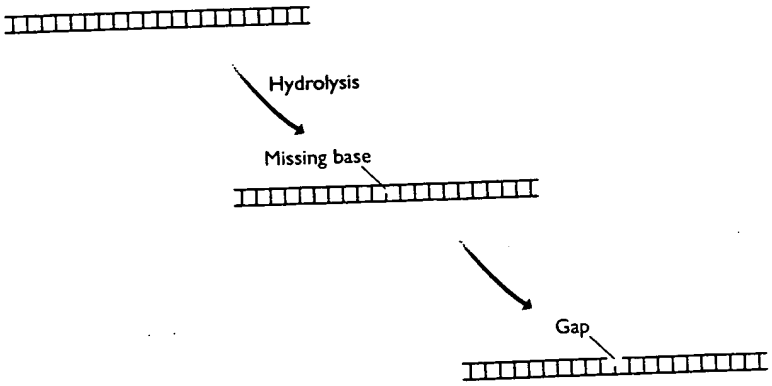


Figure 13.10 The mutagenic effect of heat.

(A) Heat induces hydrolysis of β-*N*-glycosidic bonds resulting in a baseless site in a polynucleotide. (B) Schematic representation of the effect of heat-induced hydrolysis on a double-stranded DNA molecule. The baseless site is unstable and degrades leaving a gap in one strand.

These **silent mutations** include virtually all of those that occur in extragenic DNA and in the noncoding components of genes and gene-related sequences. In other words, some 97% of the human genome (see Box 6.4, p. 135) can be mutated without significant effect.

Mutations in the coding regions of genes are much more important. First, we will look at point mutations that change the sequence of a triplet codon. A mutation of this type will have one of four effects (*Figure 13.11*):

- It may result in a **synonymous** change, the new codon specifying the same amino acid as the unmutated codon. A synonymous change is therefore a silent mutation because it has no effect on the coding function of the genome: the mutated gene codes for exactly the same protein as the unmutated gene.
- It may result in a **nonsynonymous** change, the mutation altering the codon so that it specifies a different amino acid. The protein coded by the mutated gene therefore has a single amino acid change, which often has no significant effect on the biological activity of the protein: most proteins can tolerate at least a few amino acid changes without noticeable effect on their ability to function in the cell, but changes to

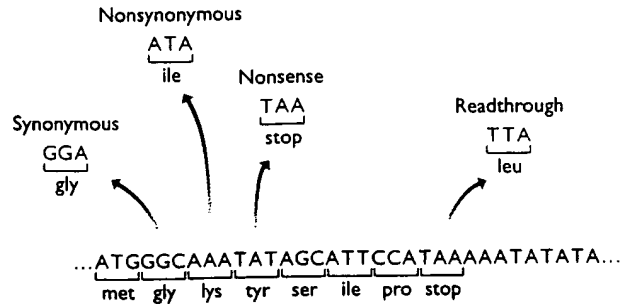


Figure 13.11 Effects of point mutations on the coding region of a gene.

Four different effects of point mutations are shown, as described in the text. The readthrough mutation results in the gene being extended beyond the end of the sequence shown here, the leucine codon created by the mutation being followed by AAA = lys, TAT = tyr and ATA = ile. See *Figure 10.7*, p. 237 for the genetic code.

some amino acids, such as those at the active site of an enzyme, have a greater impact. A nonsynonymous change is also called a **missense** mutation.

The mutation may convert a codon that specifies an amino acid into a termination codon. This is a **nonsense** mutation and it results in a shortened protein because translation of the mRNA stops at this new termination codon rather than proceeding to the correct termination codon which is further downstream. The effect that this has on the protein activity depends on how much of the polypeptide is lost; usually the effect is drastic and the protein is nonfunctional.

The mutation could convert a termination codon into one specifying an amino acid, resulting in **readthrough** of the stop signal so the protein is extended by an additional series of amino acids at its C-terminus. Most proteins can tolerate short extensions without an effect on function, but longer extensions might interfere with folding of the protein and so result in reduced activity.

Deletion and insertion mutations also have distinct effects on the coding capabilities of genes (*Figure 13.12*). If the number of deleted or inserted nucleotides is three or a multiple of three then one or more codons are removed or added, the resulting loss or gain of amino acids having varying effects on the function of the encoded protein. Deletions or insertions of this type are often inconsequential but will have an impact if, for example, amino acids involved in an enzyme's active site are lost, or if an insertion disrupts an important secondary structure in the protein. On the other hand, if the number of deleted or inserted nucleotides is not three or a multiple of three then a **frameshift** results, all of the codons downstream of the mutation being taken from a different reading frame from that used in the unmutated gene. This usually has a significant effect on the protein function, because a greater or lesser part of the mutated polypeptide has a completely different sequence to the normal polypeptide.

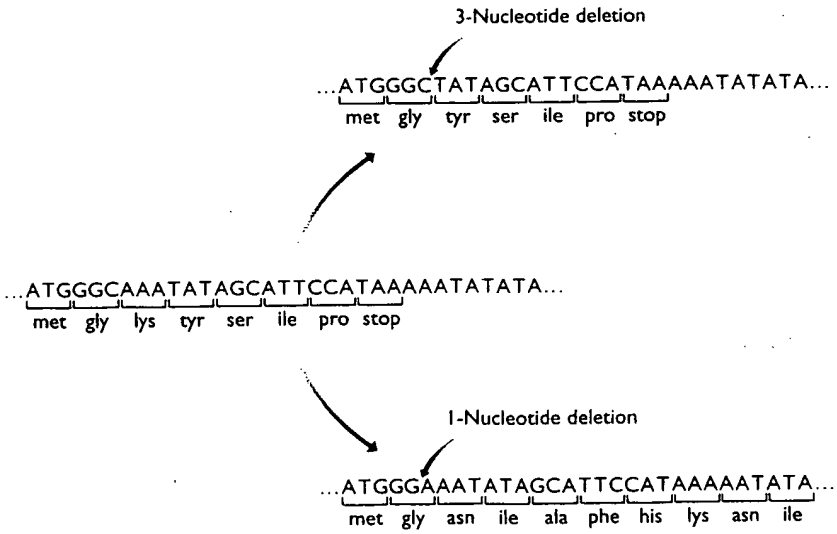


Figure 13.12 Deletion mutations.

In the top sequence three nucleotides comprising a single codon are deleted. This shortens the resulting protein product by one amino acid but does not affect the rest of its sequence. In the lower section, a single nucleotide is deleted. This results in a frameshift so all the codons downstream of the deletion are changed, including the termination codon which is now read through. See *Figure 10.7*, p. 237, for the genetic code. Note that if a three-nucleotide deletion removes parts of adjacent nucleotides then the result is more complicated than shown here. Consider, for example, deletion of the trinucleotide GCA from the sequence ...ATGGGCAAATAT... coding for Met–Gly–Lys–Tyr. The new sequence is ...ATGGAATAT..., coding for Met–Glu–Tyr. Two amino acids have been replaced by a single, different one.

It is less easy to make generalizations about the effects of mutations that occur outside of the coding regions of the genome. Any protein binding site is susceptible to point, insertion or deletion mutations that change the identity or relative positioning of nucleotides involved in the DNA–protein interaction. These mutations therefore have the potential to inactivate promoters or regulatory sequences, with predictable consequences for gene expression (*Figure 13.13*; Sections 8.2 and 8.3). Origins of replication could conceivably be made nonfunctional by mutations that change, delete or disrupt sequences recognized by the relevant binding proteins (Section 12.3.1) but these possibilities are not well-documented. There is also little information about the potential impact on gene expression of mutations that affect nucleosome positioning (Section 8.1.1).

One area that has been better researched concerns mutations that occur in introns or at intron–exon boundaries. In these regions, single point mutations will be important if they change nucleotides involved in the RNA–protein and RNA–RNA interactions that occur during splicing of different types of intron (Sections 9.2.3 and 9.3.3). For example, mutation of either the G or T in the DNA copy of the 5′ splice site of a GU–AG intron, or of the A or G at the 3′ splice site, will disrupt splicing because the correct intron–exon boundary will no longer be recognized. This may mean that the intron is not removed from the pre-mRNA, but it is more likely that a cryptic splice site (see p. 215) will be used as an alternative. It is also possible for a mutation within an intron or an exon to create a new cryptic site that is preferred over a genuine splice site that is not itself mutated. Both types of event have the same result, relocation of the active splice site, leading to aberrant splicing. This might delete or add new amino acids into the resulting protein, or lead to a frameshift. Several versions of the blood disease β-thalassemia are caused by mutations that lead to cryptic splice site selection during processing of β-globin transcripts.

The effects of mutations on multicellular organisms

Now we turn to the indirect effects that mutations have on organisms, beginning with multicellular, diploid eukaryotes such as humans. The first issue to consider is the relative importance of the same mutation in a somatic cell compared with a germ cell. Because somatic cells do not pass copies of their genomes to the next generation, a somatic cell mutation is important only for the organism in which it occurs: it has no potential evolutionary impact. In fact, most somatic cell mutations have no significant effect, even if they result in cell death, because there are many other identical cells in the same tissue and the loss of one cell is immaterial. An exception is when a mutation causes a somatic cell to malfunction in a way that is harmful to the organism, for instance by inducing tumor formation or other cancerous activity.

Mutations in germ cells are more important because they can be transmitted to members of the next generation and will then be present in all the cells of any individual who inherits the mutation. Most mutations, including all silent ones as well as many in coding regions, will still not change the phenotype of the organism in any significant way. Those that do have an effect can be divided into two categories:

- **Loss-of-function** is the normal result of a mutation that reduces or abolishes a protein activity. Most loss-of-function mutations are recessive, because in a heterozygote the second chromosome copy carries an unmutated version of the gene coding for a fully functional protein whose presence compensates for the effect of the mutation (*Figure 13.14*). There are some exceptions where a loss-of-function mutation is dominant, one example being **haploinsufficiency**, where the organism is unable to tolerate the approximately 50% reduction in protein activity suffered by the heterozygote. In humans this is the explanation of a few genetic diseases, including Marfan syndrome which results from a mutation in the gene for the connective tissue protein called fibrillin.

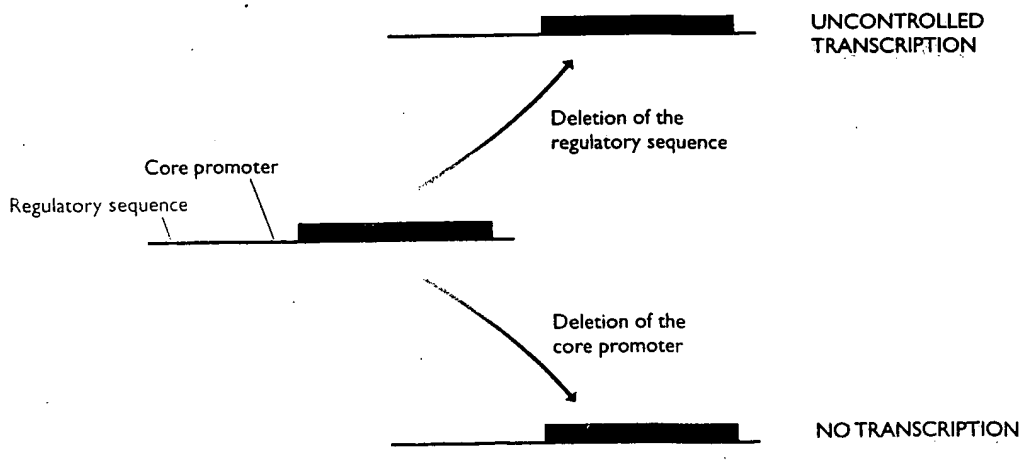


Figure 13.13 Two possible effects of deletion mutations in the region upstream of a gene.

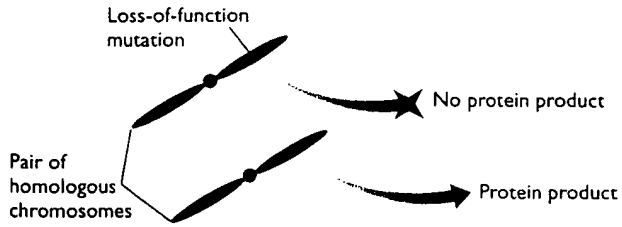


Figure 13.14 A loss-of-function mutation is usually recessive because a functional version of the gene is present on the second chromosome copy.

- **Gain-of-function** mutations are much less common. The mutation must be one that confers an abnormal activity on a protein. Many gain-of-function mutations are in regulatory sequences rather than coding regions, and can therefore have a number of consequences. For example, a mutation might lead to one or more genes being expressed in the wrong tissues, these tissues gaining functions that they normally lack. Alternatively the mutation could lead to overexpression of one or more genes involved in control of the cell cycle, thus leading to uncontrolled cell division and hence to cancer. Because of their nature, gain-of-function mutations are usually dominant.

There are added complications when considering the effects of mutations on the phenotypes of multicellular organisms. Not all mutations have an immediate effect on the organism: some are **delayed-onset** and only confer an altered phenotype later in the individual's life. Others display **nonpenetrance** in some individuals, never being expressed even though the individual has a dominant mutation or is a homozygous recessive. With humans, these factors complicate attempts to map disease-causing mutations by pedigree analysis (Section 2.3.2), because they introduce uncertainties regarding which members of a pedigree carry a mutant allele.

The effects of mutations on microorganisms

Mutations in microbes such as bacteria and yeast can also be described as loss-of-function or gain-of-function, but with microorganisms this is neither the normal nor most useful classification scheme. Instead, a more detailed description of the phenotype is usually attempted, based on the growth properties of mutated cells in various culture media. This enables most mutations to be assigned to one of four categories:

- **Auxotrophs** are cells that will only grow when provided with a nutrient not required by the unmutated organism. For example, *E. coli* normally makes its own tryptophan, courtesy of the enzymes coded by the five genes in the tryptophan operon (*Figure 6.14B*, p. 133). If one of these genes is mutated in such a way that its protein product is inactivated, then the cell is no longer able to make tryptophan and so is a tryptophan auxotroph. It cannot survive on a medium that lacks tryptophan, being able to grow only when this amino acid is provided as a nutrient (*Figure 13.15*). Unmutated bacteria, which do not require extra supplements in their growth media, are called **prototrophs**.
- **Conditional-lethal** mutants are unable to withstand certain growth conditions: under **permissive conditions** they appear to be entirely normal but when transferred to **restrictive conditions** the mutant phenotype is seen. **Temperature-sensitive** mutants are typical examples of conditional-lethals. Temperature-sensitive mutants behave like wild-type cells at low temperatures but exhibit their mutant phenotype when the temperature is raised above a certain threshold value, which is different for each mutant. Usually this is because the mutation reduces the stability of a protein, so the protein becomes unfolded and hence inactive when the temperature is raised.
- **Inhibitor-resistant** mutants are able to resist the toxic effects of an antibiotic or other type of inhibitor. There are various molecular explanations for this type of mutant. In some cases the mutation changes the structure of the protein that is targeted by the inhibitor, so the latter can no longer bind to the protein and interfere with its function. This is the basis of streptomycin-resistance in *E. coli*, which results

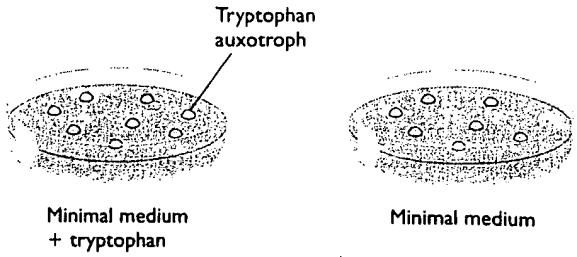


Figure 13.15 A tryptophan auxotrophic mutant.

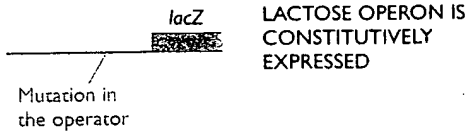
Two Petri-dish cultures are shown. Both contain **minimal medium**, which contains just the basic nutritional requirements for bacterial growth (nitrogen, carbon and energy sources, plus some salts). The medium on the left is supplemented with tryptophan but the medium on the right is not. Unmutated bacteria, plus tryptophan auxotrophs, can grow on the plate on the left, the auxotrophs growing because the medium supplies the tryptophan that they cannot make themselves. Tryptophan auxotrophs cannot grow on the plate on the right, because this does not contain tryptophan. To identify a tryptophan auxotroph, colonies are first grown on the minimal medium + tryptophan plate and then transferred to the minimal medium plate by **replica plating**. In this procedure, a sterile felt pad is pressed onto the colonies on the minimal medium + tryptophan plate, carefully removed, and then pressed onto the surface of the minimal medium plate, transferring a few bacteria from one plate to the other. After incubation, colonies appear on the minimal medium plate in the same relative positions as on the plate containing tryptophan, except for the tryptophan auxotrophs which do not grow. These colonies can therefore be identified and samples of the tryptophan auxotrophic bacteria recovered from the minimal medium + tryptophan plate.

from a change in the structure of ribosomal protein S12. Another possibility is that the mutation changes the properties of a protein responsible for transporting the inhibitor into the cell, this often being the way in which resistance to toxic metals is acquired.

Regulatory mutants have defects in promoters and other regulatory sequences. This category includes **constitutive** mutants, which continually express genes that are normally switched on and off under different conditions. For example, a mutation in the operator sequence of the lactose operon (Section 8.3.1) can prevent the repressor from binding and so results in the lactose operon being expressed all the time, even when lactose is absent and the genes should be switched off (*Figure 13.16*).

In addition to these four categories, many mutations are lethal and so result in death of the mutant cell, and others have no effect. The latter are less common in microorganisms than in higher eukaryotes, because most microbial genomes are relatively compact with little noncoding DNA. Mutations can also be **leaky**, meaning that a less extreme form of the mutant phenotype is expressed. For example, a leaky version of the tryptophan auxotroph illustrated in *Figure 13.15* would grow slowly on minimal medium, rather than not growing at all.

Is it possible for cells to utilize mutations in a positive fashion, either by increasing the rate at which mutations appear in their genomes, or by directing mutations towards specific genes? Both types of event might appear, at first glance, to go against the accepted wisdom that mutations occur randomly but, as we shall see, **hypermutation** and **programmed mutations** are possible without contravening this dogma.



Effect of a constitutive mutation in the lactose operator.

The operator sequence has been altered by a mutation and the lactose repressor can no longer bind to it. The result is that the lactose operon is transcribed all the time, even when lactose is absent from the medium. This is not the only way in which a constitutive *lac* mutant can arise. For example, the mutation could be in the gene coding for the lactose repressor, changing the tertiary structure of the repressor protein so that its DNA-binding motif is disrupted and it can no longer recognize the operator sequence, even when the latter is unmutated. See *Figure 8.15*, p. 187, for more details about the lactose repressor and its regulatory effect on expression of the lactose operon.

Hypermutation occurs when a cell causes the rate at which mutations occur in its genome to increase. This might appear to be an illogical thing to do as it is difficult to imagine situations where an increased mutation rate would be beneficial, and with the best studied example of hypermutation we do in fact know rather more about the process itself than about the reasons why it occurs. This is the **SOS response** of *E. coli*, which is induced when the genome of the bacterium suffers extensive damage, typically as a result of exposure to UV radiation or chemical mutagens. The SOS response enables the cell to replicate its DNA even though the template polynucleotides contain AP sites and/or cyclobutyl dimers and other photoproducts that would normally block or at least delay the replication complex. This requires construction of a **mutasome**, comprising several copies of the RecA protein and of the $UmuD'_2C$ complex, the latter a trimer made up of two UmuD' proteins and one copy of UmuC (Goodman, 1998). The RecA proteins coat the DNA in the region adjacent to the damage position and the $UmuD'_2C$ complexes bind to the attached RecA proteins. Somehow this enables DNA polymerase III to proceed past the damaged site and continue replicating the DNA.

The SOS response is primarily looked on as the last best chance that the bacterium has to replicate its DNA and hence survive under adverse conditions. However, the price of survival is an increased mutation rate because the mutasome does not repair damage, it simply allows a damaged region of a polynucleotide to be replicated. When it encounters a damaged position in the template DNA the polymerase selects a nucleotide more or less at random, though with some preference for placing an A opposite an AP site: in effect the error rate of the replication process becomes increased. It has been suggested that this increased mutation rate is the purpose of the SOS response, mutation for some reason or other being an advantageous response to DNA damage, but this idea remains controversial (Walker, 1995).

Less controversial is the way in which vertebrates, including humans, are able to increase the mutation rate at one specific gene in one type of cell. This phenomenon takes us back to the way in which immunoglobulin diversity is generated, which we have already touched upon in Section 11.2.1 when we examined the genome rearrangements that result in joining of the V, D, J and H segments of the immunoglobulin heavy and light genes (see *Figure 11.14*, p. 280). Additional diversity is produced by hypermutation of the V gene segments, after assembly of the intact immunoglobulin gene (*Figure 13.17*), the mutation rate for these segments being 6–7 orders of magnitude greater than the background mutation rate experienced by the rest of the genome (Shannon and Weigert, 1998). This enhanced mutation rate appears to result from the unusual behavior of the mismatch repair system which normally corrects replication errors. At all other positions within the genome, the mismatch repair system corrects

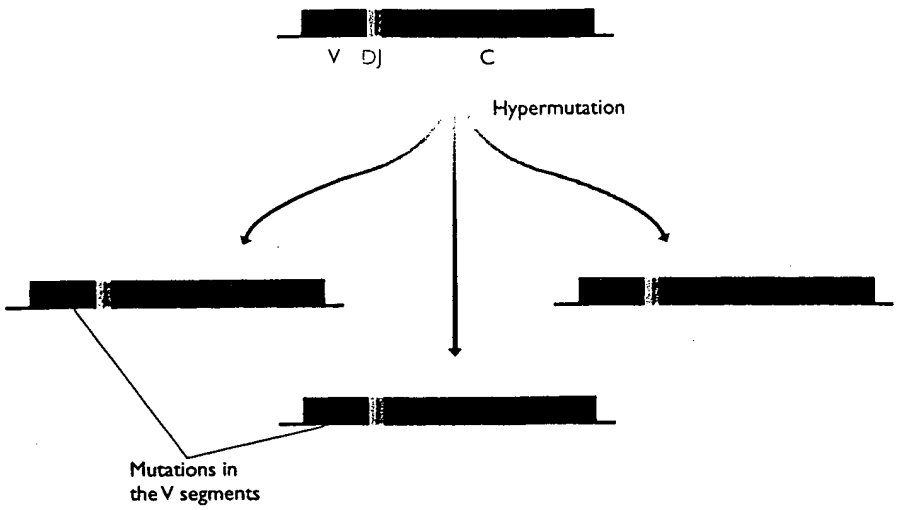


Figure 13.17 Hypermutation of the V gene segment of an intact immunoglobulin gene.

See *Figure 11.14*, p. 280, for a description of the events leading to assembly of an immunoglobulin gene.

errors of replication by searching for mismatches and replacing the nucleotide in the daughter strand, this being the strand that has just been synthesized and so contains the error (see Section 13.1.4). At V gene segments, the repair system changes the nucleotide in the parent strand, and so stabilizes the mutation rather than correcting it (Cascalho *et al.*, 1998). The mechanism by which this is achieved has not yet been described.

An apparent increase in mutation rate arising from modifications to the normal DNA repair process does not contradict the dogma regarding the randomness of mutations. Where problems have arisen is with reports, dating back to 1988 (Cairns *et al.*, 1988), suggesting that *E. coli* is able to direct mutations towards genes whose mutation would be advantageous under the environmental conditions that the bacterium is encountering. The original experiments involved a strain of *E. coli* that has a frameshift mutation in the lactose operon, inactivating the proteins needed for utilization of this sugar (Research Briefing 13.1). The bacteria were spread on an agar medium in which the only carbon source was lactose. This meant that a cell could grow and divide only if a second mutation occurred in the lactose operon, restoring the correct reading frame and therefore allowing the lactose enzymes to be synthesized. Mutations with this effect appeared to occur significantly more frequently than expected, and at a rate that was greater than mutations in other parts of the genomes of these *E. coli* cells.

These experiments suggested that bacteria can programme mutations according to the selective pressures that they are placed under. In other words, the environment can directly affect the phenotype of the organism, as suggested by Lamarck, rather than operating through the random processes postulated by Darwin. With the implications being so radical it is not surprising that the experiments have been debated at length with numerous attempts to discover flaws in their design or alternative explanations for the results. At present, the possibility that the enhanced mutation rate is not programmed towards specific genes but occurs throughout the genome is being re-examined (Bridges, 1997), and models based on gene amplification rather than selective mutation are being tested (Andersson *et al.*, 1998).

13.1.4 DNA repair

In view of the thousands of damage events that genomes suffer every day, coupled with the errors that occur when the genome replicates, it is essential that cells possess efficient repair systems. Without these repair systems a genome would not be able to maintain its essential cellular functions for more that a few hours before key genes became inactivated by DNA damage. Similarly, cell lineages would accumulate replication errors at such a rate that their genomes would become dysfunctional after a few cell divisions.

Most cells possess five different categories of DNA repair system:

- **Direct repair systems**, as the name suggests, act directly on damaged nucleotides, converting each one back to its original structure.
- **Base excision repair** involves removal of a damaged nucleotide base, excision of a short piece of the polynucleotide around the AP site thus created, and resynthesis with a DNA polymerase.
- **Nucleotide excision repair** is similar to base excision repair but is not preceded by removal of a damaged base and can act on more substantially damaged areas of DNA.
- **Mismatch repair** corrects errors of replication, again by excising a stretch of single-stranded DNA containing the offending nucleotide and then repairing the resulting gap.

RESEARCH

13.1

BRIEFING

Adaptive mutations?

In 1988 startling results were published suggesting that under some circumstances *Escherichia coli* bacteria are able to mutate in a directed way that enables cells to adapt to an environmental stress.

The randomness of mutations is an important concept in biology because it is a requirement of the Darwinian view of evolution, which holds that changes in the characteristics of an organism occur by chance and are not influenced by the environment in which the organism is placed. Beneficial changes are positively selected and harmful ones are negatively selected (see Box 15.5, p. 408). In contrast, the Lamarckian theory of evolution, which biologists rejected well over a century ago, states that organisms acquire changes that enable them to adapt to their environment. The Darwinian view requires that mutations occur at random, whereas Lamarckian evolution demands that adaptive mutations occur in response to the environment.

Random mutations in E. coli

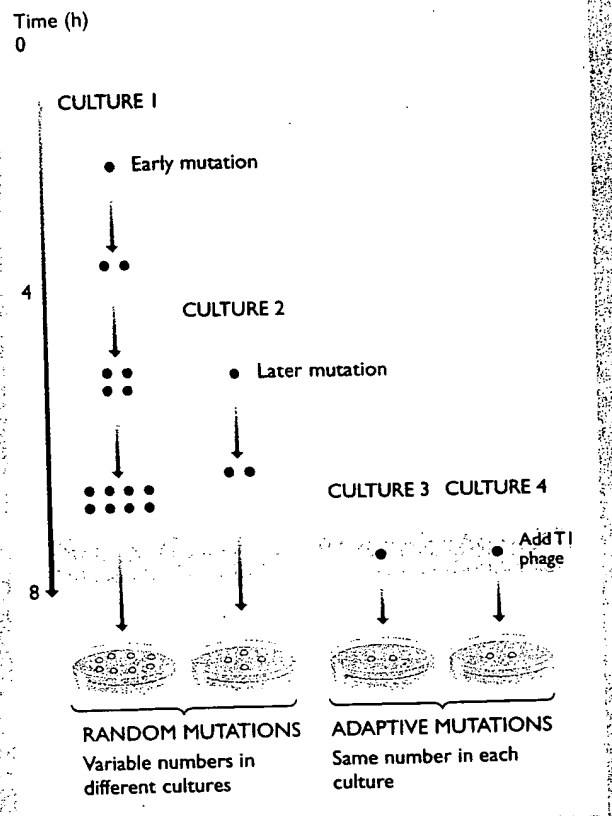
The randomness of mutations in bacteria was first demonstrated by Luria and Delbrück in 1943. They grew a series of *E. coli* cultures in different flasks and then added T1 bacteriophages to each one. Most of the bacteria were killed by the phage, but a few T1-resistant mutants were able to survive. These were identified by plating samples from each culture, soon after T1 infection, onto an agar medium. If mutations leading to T1 resistance occurred randomly in the cultures before the bacteriophages were added, each culture would contain different numbers of resistant mutants, the numbers depending on how early during the growth period the first mutant cells arose. Those that arose early would divide many times to give rise to a large number of resistant progeny in the culture at the end of the growth period, whereas those that arose later would give rise to just a few progeny. Some cultures would therefore contain many T1-resistant cells and others would contain just a few. Alternatively, if resistant bacteria arose by adaptive mutation only when the T1 phage were added, then all cultures would have similar numbers of mutants (see figure opposite).

Luria and Delbrück found that each of their cultures contained a different number of T1-resistant bacteria; thus, they concluded that mutations occur randomly and not in response to T1 phage.

Adaptive mutations in E. coli

The possibility that Luria and Delbrück's conclusion might not be universally true for *E. coli* mutations was first suggested by studies of an *E. coli* strain that carries a nonsense mutation in its *lacZ* gene. The presence of the termination codon in *lacZ* means that these cells are unable to synthesize functional β-galactosidase enzymes and so cannot use lactose as a carbon and energy source – they are therefore lactose auxotrophs. This is not necessarily a permanent situation because a cell could undergo a mutation that converts the termination codon back into one specifying an amino acid. These new mutants would be able to make β-galactosidase and use any lactose that is available. According to Luria and Delbrück's results, such mutations should occur at random and should not be influenced by the presence of lactose in the medium. The results of Cairns *et al.* (1988) showed that when the lactose auxotrophs were plated onto a minimal medium containing lactose as the only sugar – circumstances that require that the bacteria must mutate into lactose prototrophs in order to survive – then the number of lactose prototrophs that arose was significantly higher than that expected if mutations occurred randomly. In other words, some cells mutated adaptively and acquired the specific DNA sequence change needed to withstand the selective pressure.

Since 1988, a number of examples of what appear to be adaptive mutations have been published, but the notion that bacteria, and possibly other organisms, can program mutations in response to environmental stress is by no means accepted by the scientific community. It is quite possible that these mutations will eventually be disproved or shown to have an orthodox basis. However, until this happens we are left with the tantalizing possibility that even at this very fundamental level our knowledge about genomes might be far from complete.



Reference

Cairns J, Overbaugh J and Miller S (1988) The origin of mutants. *Nature*, **335**, 142–145.

Box 13.2: DNA repair and human disease

The importance of DNA repair is emphasized by the number and severity of inherited human diseases which have been linked with defects in one of the repair processes. One of the best characterized of these is xeroderma pigmentosum, which results from a mutation in any one of several genes for proteins involved in nucleotide excision repair. The disease symptoms include hypersensitivity to UV radiation, so patients suffer more mutations on exposure to sunlight, this often leading to skin cancer (Lehmann, 1995). Two other diseases, Cockayne syndrome and trichothiodystrophy, are also caused by defects in nucleotide excision repair, but these are more complex disorders which, although not involving cancer, usually include problems with both the skin and nervous system.

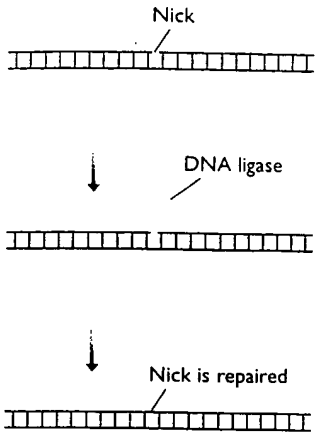
A few diseases have been linked with defects in the transcription-coupled component of nucleotide excison repair. These include breast and ovarian cancers, the BRCA1 gene that confers susceptibility to these cancers coding for a protein that has been implicated, at least indirectly, with transcription-coupled repair (Gowen *et al.*, 1998). A deficiency in transcription-coupled repair has also been identified in humans suffering from the cancer susceptibility syndrome called HNPCC (hereditary nonpolyposis colorectal cancer; Mellon *et al.*, 1996) though this disease was originally identified as a defect in mismatch repair (Kolodner, 1995). Other diseases that seem to involve a breakdown in some aspect of DNA repair, but whose direct causes have not been uncovered, are ataxia telangiectasia, whose symptoms include sensitivity to ionizing radiation, Bloom's syndrome, which is probably due to inactivation of DNA ligase genes, and Fanconi's anemia, which confers sensitivity to chemicals that cause crosslinks in DNA.

Recombination repair is used to mend double-strand breaks.

In this section we will look at the first four types of repair system, leaving recombination repair until Box 13.4, this last system being easier to understand after we have dealt with the more general principles of recombination.

Relatively few forms of DNA damage can be repaired without excision of nucleotides. Those that can be repaired by direct methods are as follows:

Nicks can be repaired by a DNA ligase if all that has happened is that a phosphodiester bond has been broken, without damage to the 5′-phosphate and 3′-hydroxyl groups of the nucleotides either side of the nick (*Figure 13.18*). This is often the case with nicks resulting from the effects of ionizing radiation.



Repair of a nick by DNA ligase.

Some forms of **alkylation** damage are directly reversible by enzymes that transfer the alkyl group from the nucleotide to their own polypeptide chains. Enzymes capable of doing this are known in many different organisms and include the **Ada enzyme** of *E. coli*, which is involved in an adaptive process that this bacterium is able to activate in response to DNA damage. Ada removes alkyl groups attached to the oxygen groups at positions 4 and 6 of thymine and guanine, respectively, and can also repair phosphodiester bonds that have become methylated. Other alkylation repair enzymes have more restricted specificities, an example being human **MGMT** (O^6-methylguanine-DNA methyltransferase) which, as its name suggests, only removes alkyl groups from position 6 of guanine.

Cyclobutyl dimers are repaired by a light-dependent direct system called **photoreactivation**. In *E. coli*, the process involves the enzyme called **DNA photolyase** (more correctly named deoxyribodipyrimidine photolyase). When stimulated by light with a wavelength between 300 and 500 nm the enzyme binds to cyclobutyl dimers and converts them back to the original monomeric nucleotides. Photoreactivation is a widespread but not universal type of repair: it is known in many but not all bacteria and also in quite a few eukaryotes, including some vertebrates, but is absent in humans. A similar type of photoreactivation involves the **(6–4) photoproduct photolyase** and results in repair of (6–4) lesions. Neither *E. coli* nor humans have this enzyme but it is possessed by a variety of other organisms.

Base excision is the least complex of the various repair systems that involve removal of a damaged nucleotide followed by resynthesis of DNA to span the resulting gap. It is used to repair many modified nucleotides that have

Box 13.3: Cell cycle checkpoints for monitoring DNA damage

In a multicellular organism, the death of a single somatic cell as a result of DNA damage is usually less dangerous than allowing that cell to replicate its mutated DNA and possibly give rise to a tumor or other cancerous growth. Eukaryotic cells therefore monitor their genomes for damage, principally at the checkpoints immediately before the entry into the S and M phases of the cell cycle (Section 12.4.1; Russell, 1998). These checkpoints ensure that a damaged genome is not replicated, which would lead to mutations being perpetuated in the cell lineage, and prevent problems with the distribution of chromosomes to daughter cells, which might occur if one or more chromosomes has extensive DNA damage. A cell that fails to pass one or other of the checkpoint tests might undergo cell cycle arrest, permanently or until its DNA is repaired, or it might be forced into programmed cell death or **apoptosis** (Chernova *et al.*, 1995; Enoch and Norbury, 1995).

In mammals, a central player in induction of cell cycle arrest and apoptosis is the protein called p53. This is classified as a tumor-suppressor protein, because when this protein is defective, cells with damaged genomes can avoid the cell cycle checkpoints and possibly proliferate into a cancer. p53 is a sequence-specific DNA-binding protein that activates a number of genes thought to be directly responsible for arrest and apoptosis, and also represses expression of others that must be switched off to facilitate these processes. A second protein that might play a regulatory role at the checkpoints is the product of the human *ATM* gene which, when defective, gives rise to ataxia telangiectasia, one of the diseases associated with a deficiency in DNA repair (see Box 13.2).

suffered relatively minor damage to their bases. The process is initiated by a **DNA glycolyase** which cleaves the β-*N*-glycosidic bond between a damaged base and the sugar component of the nucleotide (*Figure 13.19A*). Each DNA glycolyase has a limited specificity, the specificities of the glycolyases possessed by a cell determining the range of damaged nucleotides that can be repaired by the base excision pathway. Most organisms are able to deal with deaminated bases such as uracil (deaminated cytosine) and hypoxanthine (deaminated adenine), oxidation products such as 5-hydroxycytosine and thymine glycol, and methylated bases such as 7-methylguanine and 2-methylcytosine (Seeberg *et al.*, 1995). Other DNA glycolyases remove normal bases as part of the mismatch repair system.

DNA glycolyase removes a damaged base by 'flipping' the structure to a position outside of the helix and then detaching it from the polynucleotide (Kunkel and Wilson, 1996; Roberts and Cheng, 1998). This creates an AP or baseless site (see *Figure 13.10*, p. 341) which is converted into a single nucleotide gap in the second step of the repair pathway (*Figure 13.19B*). This step can be carried out in a variety of ways. The standard method makes use of an **AP endonuclease**, such as exonuclease III or endonuclease IV of *E. coli*, which cuts the phosphodiester bond on the 5′ side of the AP site. Some AP endonucleases can also remove the sugar from the AP site, this being all that remains of the damaged nucleotide, but others lack this ability and so work in conjunction with a separate **phosphodiesterase**. An alternative pathway for converting the AP site into a gap utilizes the endonuclease activity possessed by some DNA glycolyases, which can make a cut at the 3′ side of the AP site, probably at the same time that the damaged base is removed, followed again by removal of the sugar by a phosphodiesterase.

The single nucleotide gap is filled by a DNA polymerase, using the undamaged base in the other strand of the DNA molecule to ensure that the correct nucleotide is inserted. In *E. coli* the gap is filled by DNA polymerase I and in mammals by DNA polymerase β (see *Table 12.2*, p. 313; Sobol *et al.*, 1996). Yeast seems to be unusual in that it uses its main DNA replicating enzyme, DNA polymerase δ, for this purpose (Seeberg *et al.*, 1995). After gap filling, the final phosphodiester bond is put in place by a DNA ligase.

Nucleotide excision repair is used to correct more extensive types of damage

Nucleotide excision repair has a much broader specificity than the base excision system and is able to deal with more extreme forms of damage such as intrastrand crosslinks and bases that have become modified by attachment of large chemical groups. It is also able to correct cyclobutyl dimers by a **dark repair** process, providing those organisms that do not have the photoreactivation system with a means of repairing these dimers.

In nucleotide excision repair, a segment of single-stranded DNA containing the damaged nucleotide(s) is excised and replaced with new DNA. The process is therefore similar to base excision repair except that it is not preceded by selective base removal and a longer stretch of polynucleotide is excised. The best studied example of nucleotide excision repair is the **short patch** process of *E. coli*, so called because the region of polynucleotide that is excised and subsequently 'patched' is relatively short, usually 12 nucleotides in length.

Short patch repair is initiated by a multienzyme complex called the **UvrABC endonuclease**, sometimes also referred to as the 'excinuclease'. In the first stage of the process a trimer comprising two UvrA proteins and one copy of UvrB attaches to the DNA at the damaged site. How the site is recognized is not known but the broad specificity of the process indicates that individual types of damage are not directly detected and that the complex must search for a more general attribute of DNA damage such as distortion of the double helix. UvrA may be the part of the complex most involved in damage location because it dissociates once the site has been found and plays no further part in the repair process. Departure of

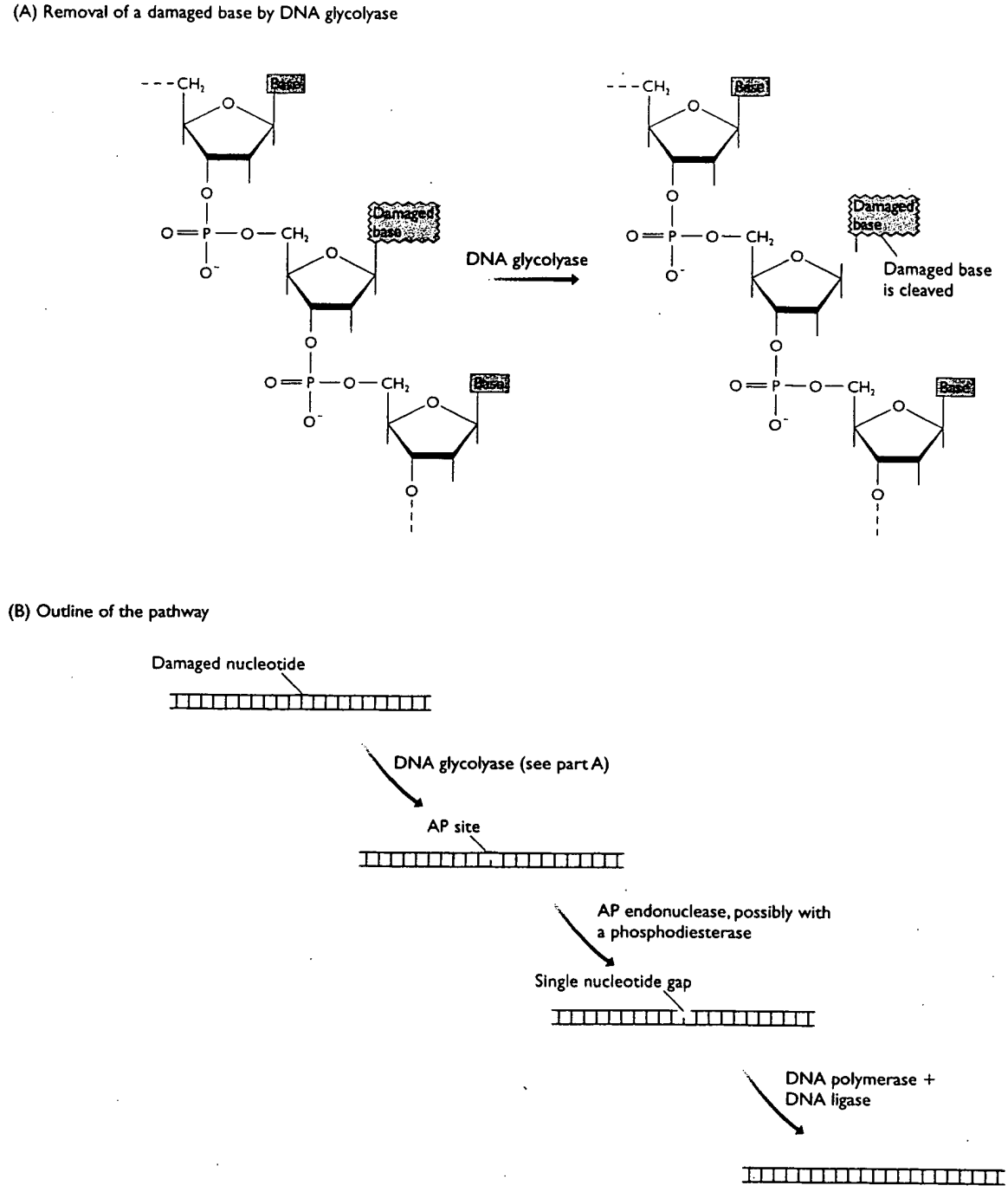
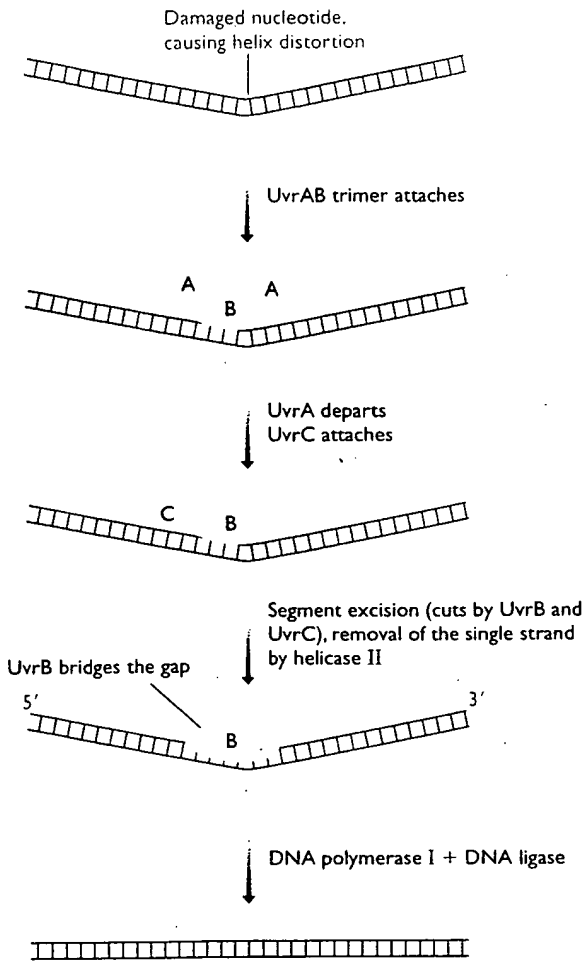


Figure 13.19 Base excision repair.

(A) Excision of a damaged nucleotide by a DNA glycolyase. (B) Schematic representation of the base excision repair pathway. Alternative versions of the pathway are described in the text.

UvrA allows UvrC to bind (*Figure 13.20*), forming a UvrBC dimer that cuts the polynucleotide either side of the damaged site. The first cut is made by UvrB at the fifth phosphodiester bond downstream of the damaged nucleotide, and the second cut is made by UvrC at the eighth phosphodiester bond upstream, resulting in the 12 nucleotide excision, though there is some variability, especially in the position of the UvrB cut site. The excised segment is then removed, usually as an intact oligonucleotide, by DNA helicase II which presumably detaches the segment by breaking the base pairs holding it to the second strand. UvrC also detaches at this stage, but UvrB remains in place and bridges the gap produced by the excision, possibly to prevent the single-stranded region



Short patch nucleotide excision repair in *E. coli*.

The damaged nucleotide is shown distorting the helix because this is thought to be one of the recognition signals for the UvrAB trimer that initiates the short patch process. See the text for details of the events occurring during the repair pathway.

that has been exposed from base-pairing with itself, possibly to prevent this strand from becoming damaged, or possibly to direct the DNA polymerase to the site that needs to be repaired. As in base excision repair, the gap is filled by DNA polymerase I and the last phosphodiester bond is synthesized by DNA ligase.

E. coli also has a **long patch** nucleotide excision repair system that involves Uvr proteins but differs in that the piece of DNA that is excised can be anything up to 2 kb in length. Long patch repair has been less well studied and the process is not understood in detail, but it is presumed to work on more extensive forms of damage, possibly regions where groups of nucleotides, rather than just single bases, have become modified. The eukaryotic nucleotide excision repair process is also called 'long patch' but results in replacement of only 24–29 nucleotides of DNA. In fact, there is no 'short patch' system in eukaryotes and the name is used to distinguish the process from base excision repair. The system is more complex than in *E. coli* and the relevant enzymes do not seem to be homologs of the Uvr proteins. In humans at least 16 proteins are involved, with the downstream cut being made at the same position as in *E. coli* – the fifth phosphodiester bond – but with a more distant upstream cut, resulting in the longer excision. Both cuts are made by endonucleases that attack single-stranded DNA specifically at its junction with a double-stranded region, indicating that before the cuts are made the DNA around the damage site has been melted, presumably by a helicase (*Figure 13.21*). This activity is provided at least in part by TFIIH, one of the components of the RNA polymerase II initiation complex (see *Table 8.3*, p. 184). At first it was assumed that TFIIH simply has a dual role in the cell, functioning separately in both transcription and repair, but now it is thought that there is a more direct link between the two processes (Lehmann, 1995; Svejstrup *et al.*, 1996). This view is supported by the discovery of **transcription-coupled repair**, which results in the template strands of genes being repaired more quickly than other parts of the eukaryotic genome, an observation that is entirely logical as these template strands contain the genome's biological information and maintaining their integrity is the highest priority for the repair systems.

Each of the three repair systems that we have looked at so far recognize and act upon DNA damage caused by mutagens. This means that they search for abnormal chemical structures such as modified nucleotides, cyclobutyl dimers and intrastrand crosslinks. They cannot correct mismatches resulting from errors in replication because the mismatched nucleotide is not abnormal in any way, it is simply an A, C, G or T that has been inserted at the wrong position. As these nucleotides look exactly like any other nucleotide the mismatch repair system that corrects replication errors has to detect not the mismatched nucleotide itself but the absence of base-pairing between the parent and daughter strands. Once it has found a mismatch, the repair system excises part of the daughter polynucleotide and fills in the gap, in a manner similar to what we have already seen with base and nucleotide excision repair.

The scheme described above leaves one important question unanswered. The repair must be made in the daughter polynucleotide because it is in this newly synthesized strand that the error has occurred: the parent polynucleotide has the correct sequence. How does the repair process know which strand is which? In *E. coli* the answer is that the daughter strand is, at this stage, undermethylated and therefore can be distinguished from the parent polynucleotide, which has a full complement of methyl groups. *E. coli* DNA is methylated because of the activities of the **DNA adenine methylase (Dam)**, which converts adenines to 6-methyladenines in the sequence 5′-GATC-3′, and the **DNA cytosine methylase**

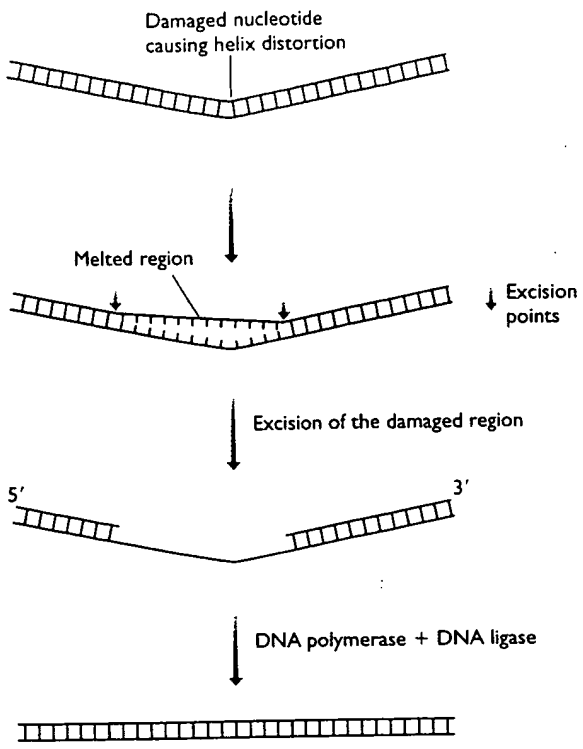


Figure 13.21 Outline of the events involved in nucleotide excision repair in eukaryotes.

The endonucleases that remove the damaged region make cuts specifically at the junction between single-stranded and double-stranded regions of a DNA molecule. The DNA is therefore thought to melt either side of the damaged nucleotide, as shown in the diagram, possibly due to the helicase activity of TFIIH.

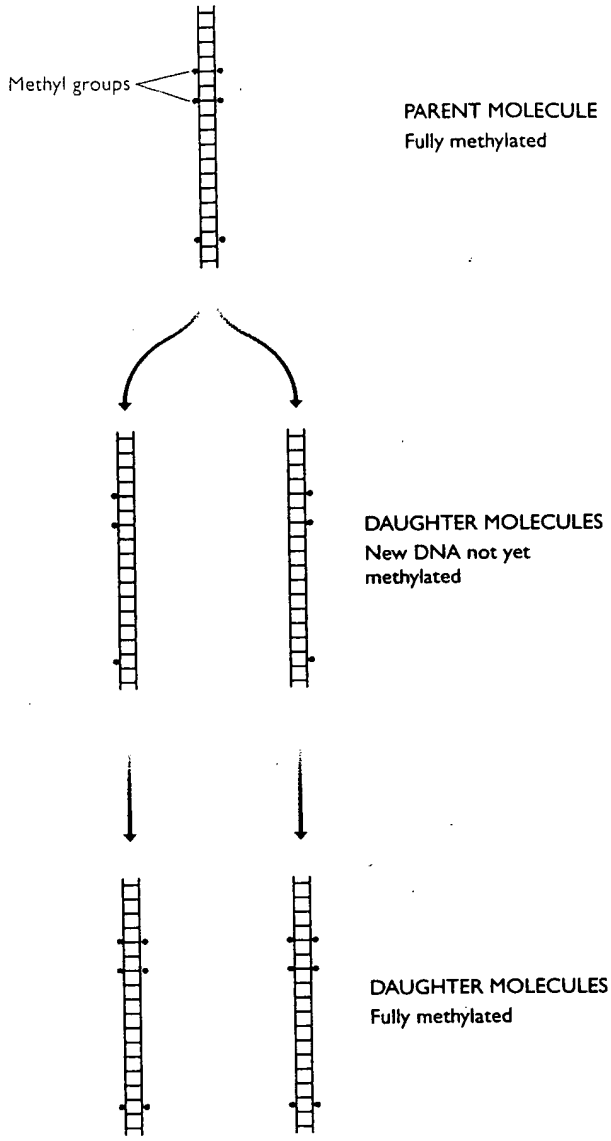
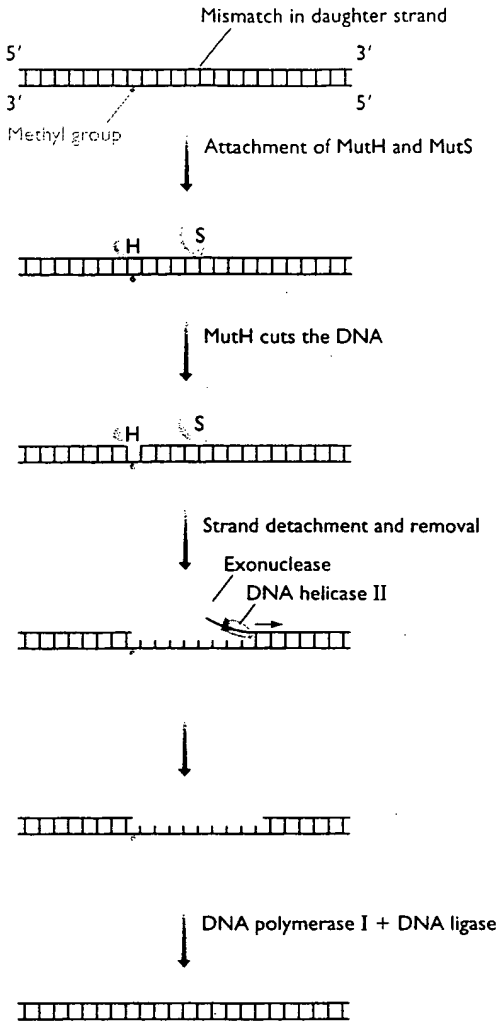


Figure 13.22 Methylation of newly-synthesized DNA in *E. coli* does not occur immediately after replication, providing a window of opportunity for the mismatch repair proteins to recognize the daughter strands and correct replication errors.

(Dcm), which converts cytosines to 5-methylcytosines in 5′-CCAGG-3′ and 5′-CCTGG-3′. These methylations are not mutagenic, the modified nucleotides having the same base-pairing properties as the unmodified versions. There is delay between DNA replication and methylation of the daughter strand, and it is during this window of opportunity that the repair system scans the DNA for mismatches and makes the required corrections in the undermethylated, daughter strand (*Figure 13.22*).

E. coli has at least three mismatch repair systems, called 'long patch', 'short patch and 'very short patch', the names indicating the relative lengths of the excised and resynthesized segments. The long patch system replaces up to a kb or more of DNA and requires the MutH, MutL and MutS proteins, as well as the DNA helicase II that we met during nucleotide excision repair. MutS recognizes the mismatch and MutH distinguishes the two strands by binding to unmethylated 5′-GATC-3′ sequences (*Figure 13.23*). The role of MutL is unclear but it might coordinate the activities of the two other proteins so that MutH binds to 5′-GATC-3′ sequences only in the vicinity of mismatch sites recognized by MutS. After binding, MutH cuts the phosphodiester bond immediately upstream of the G in the methylation sequence and DNA helicase II detaches the single strand. There does not appear to be an enzyme that cuts the strand downstream of the mismatch; instead the detached single-stranded region is degraded by an exonuclease that follows the helicase and continues beyond the mismatch site. The gap is then filled in by DNA polymerase I and DNA ligase.

Similar events are thought to occur during short and very short mismatch repair, the difference being the specificities of the proteins that recognize the mismatch. The short patch system, which results in excision of a segment



Long patch mismatch repair in *E. coli*.

See the text for details.

less than 10 nucleotides in length, begins when MutY recognizes an A–G or A–C mismatch, and the very short repair system corrects G–T mismatches which are recognized by the Vsr endonuclease.

Eukaryotes have homologs of the *E. coli* Mut proteins and their mismatch repair processes probably work in a similar way. The one difference is that methylation might not be the method used to distinguish between the parent and daughter polynucleotides (Modrich and Lahue, 1996). Methylation has been implicated in mismatch repair in mammalian cells, but the DNA of some eukaryotes, including fruit flies and yeast, is not extensively methylated (see Box 8.1, p. 177); it is thought that these organisms must therefore use a different method. Possibilities include an association between the repair enzymes and the replication complex, so that repair is coupled with DNA synthesis, or use of single-strand binding proteins that mark the parent strand.

13.2 RECOMBINATION

Without recombination, genomes would be relatively static structures, undergoing very little change. Over a long period of time the gradual accumulation of mutations would result in small scale alterations in the nucleotide sequence of the genome, but more extensive restructuring, which is the role of recombination, would not occur. The evolutionary potential of the genome would be severely restricted.

Recombination was first recognized as the process responsible for crossing-over and exchange of DNA segments between homologous chromosomes during meiosis of eukaryotic cells (see *Figure 2.10*, p. 26), and subsequently implicated in the integration of transferred DNA into bacterial genomes after conjugation, transduction or transformation (Section 2.3.2). The biological importance of these processes stimulated the first attempts to describe the molecular events involved in recombination and led to the Holliday model (Holliday, 1964), with which we will begin our study of recombination.

The Holliday model refers to a type of recombination called **general** or **homologous recombination**. This is the most important version of recombination in nature, being responsible for meiotic crossing-over and integration of transferred DNA into bacterial genomes.

The Holliday model describes recombination between two homologous double-stranded molecules, ones with identical or nearly identical sequences, but is equally applicable to two different molecules that share a limited region of homology, or a single molecule that recombines with itself because it contains two separate regions that are homologous with one another.

The central feature of the model is formation of a **heteroduplex** resulting from the exchange of polynucleotide segments between the two homologous molecules (*Figure 13.24*). The heteroduplex is initially stabilized by base-pairing between each transferred strand and the intact polynucleotide of the recipient molecule, this base-pairing being possible because of the sequence similarity between the two molecules. Subsequently the gaps are sealed by DNA ligase, giving a **Holliday structure**. This structure is dynamic, **branch migration** resulting in exchange of longer segments of DNA being possible if the two helices rotate in the same direction.

Separation, or **resolution**, of the Holliday structure back into individual double-stranded molecules occurs by cleavage across the branchpoint. This is the key to the entire process because the cut can be made in either of two orientations, as becomes apparent when the three dimensional configuration or **chi form** of the Holliday structure is examined (see *Figure 13.24*). These two cuts have very different results. If the cut is made left–right